CONTEMPORARY PRINCIPLES TO CONTROL COMPLEX OBJECTS

L. A. RASTRIGIN

MIR PUBLISHERS MOSCOW

This book is the first in the world to attempt a 'thorough' presentation of the contemporary principles, approaches, and methods to control complex, but formal, objects of diverse physical natures.

The book is an intermediate one between a scientific monograph and popular science. It introduces the reader to the problem of controlling complex objects and is in a form accessible to an engineer.

MIR PUBLISHERS of Moscow publishes Soviet scientific and technical literature in nineteen languages—Russian, English, German, French, Italian, Spanish, Portuguese, Czech, Slovak, Finnish, Hungarian, Mongolian, Arabic, Persian, Hindi, Vietnamese, Tamil, Kannada, and Dari. Titles include textbooks for higher technical schools and vocational schools, literature on the natural sciences and medicine, including textbooks for medical schools, popular science and science fiction. The contributors to MIR PUBLISHERS' list are leading Soviet scientists and engineers in all fields of science and technology and include more than 40 Members and Corresponding Members of the USSR Academy of Sciences. Skilled translators provide a high standard of translation from the original Russian.

Many of the titles already issued by MIR PUBLISHERS have been adopted as textbooks and manuals at educational establishments in France, Switzerland, Cuba, Syria, India, Brasil and many other countries.

MIR PUBLISHERS' books in foreign languages are exported [illegible] narodnaya Kniga" [illegible] ordered through [illegible] try dealing with [illegible] Kniga".

Professor Rastrigin is the author of ten scientific monographs in the field of technical cybernetics. His professional interests are related to random search which is a new and efficient method for solving the complex problems of control. The author's hobby is popularization of his science and he has written seven popular science books which have been translated into all the main European and Oriental languages.

Graduating in 1953 from the Moscow Aeronautical Institute, the author received the degree of Doctor of Tech. Sc. in the field of statistical optimization in 1965. He is currently the head of the 'Adaptive systems' laboratory at the Institute of Electronics and Computers, Latvian Acad. Sc., Riga.

CONTEMPORARY PRINCIPLES TO CONTROL COMPLEX OBJECTS

L.A. RASTRIGIN

CONTEMPORARY PRINCIPLES TO CONTROL COMPLEX OBJECTS

Л. А. Растригин

Современные принципы управления сложными объектами

Издательство «Советское радио»
Москва

CONTEMPORARY PRINCIPLES TO CONTROL COMPLEX OBJECTS

L. A. RASTRIGIN

Translated from the Russian
by Michael Burov

Mir Publishers Moscow

First published 1983
Revised from the 1980 Russian edition

На английском языке

© Издательство «Советское радио», 1980
© English translation, Mir Publishers, 1983

From the Author

Knowing some principles quite commonly makes up for not knowing some facts.

Helvetius

A tendency to use formal (mathematical) methods to control complex objects is characteristic for our time. A great deal has been done in this way, but the use of informal (expert) methods in the synthesis of a process to control a complex object is also unavoidable. This should not be regarded as 'sweeping the dust under the rug' since any expert is as good a source of information as a sensor, though just like any sensor he has features of his own, that need to be accounted for. It is a peculiarity tackled in the book that the control of a complex object requires a combination of both formal and informal methods.

All the stages of the control of a complex object are considered from the determination of the control objectives to the correction of these objectives as the created control system operates.

The process of achieving control is considered in the book from the point of view of algorithms, irrespective of the computer hardware for implementing the algorithms (or whether they are going to be implemented at all). There are special books on the design of control systems which deal with the ways and means to implement the algorithms with the aid of computers, see, for instance, Zakharov, V. N. et al., *Control Systems*, Energiya, Moscow, 1977 (in Russian).

The aim of the author is to discuss each of the eight stages of the control of a complex object at the level needed by an engineer. Each chapter analyzes a stage in the creation of a control system. Thus, Chapter 1 deals with the first stage of any control, the determination of the objectives of control, and Chapter 2 discusses the second stage, isolation of the object of control from the medium. The next three chapters deal with synthesis of the specified control model. Chapter 3 deals with the structural synthesis of the model, and Chapters 4 and

5 with the parametric, passive synthesis (for identification) and the active synthesis (for experimentation with the object). The synthesis of a control system on the basis of the model of the object that has been produced is considered in Chapter 6. Chapter 7 deals with the stage at which control is implemented. And finally, the last chapter discusses the stage when the control system (the model, the object, and the objectives) is corrected on the grounds of the data obtained in the process of control.

The book presents original material as well, for instance, all of the first two chapters cover the author's research in determining the objectives and isolating the object from the medium. The other chapters blend the author's concepts with what is known to professionals in the field.

Naturally, it is impossible to embrace in just eight chapters the 'unembraceable' problem of the control of complex objects. Therefore, the author considers only those algorithms, methods, and approaches of the theory and practice of control which best reflect the contemporary principles of complex object control. It is quite evident that some aspects of the problem of control were not analyzed at all: there is no mention of the hierarchical approach to description of complex objects and their control, the so-called systems approach, and the models of automatized control systems are disregarded as well, etc. The main reason is the limited space in the book.

It would be an error to consider the book a reference manual, its aim is to give the reader an idea of the principles, approaches and methods of controlling complex objects now in effective use. The recipes necessary to apply the matter presented may be found in the references, comments on which are supplied at the end of each chapter.

The book is based on an undergraduate course of lectures given by the author for more than ten years now at the department of automatics and computers at the Riga Polytechnical Institute. This has predetermined the synthetic and generalizing nature of presenting the material which is supported by student's lore formed during all the years of study. This has also made possible (and compelling) to consider almost all the sundry methods of

contemporary cybernetics from the uniform viewpoint of control of a complex object.

An undergraduate course in mathematics is sufficient for the contents of the book to be comprehended but the special problems of optimization are discussed in the 'Supplement'. For a superficial understanding of the basic ideas and principles, only the Introduction (for which mathematics is not necessary) need be read. If the reader goes past the Introduction, the author's goal will have been achieved.

The manuscript was reviewed by V. A. Abchuk and I. A. Poletayev who made numerous critical remarks, which caused the author to revise the book considerably. It was also read by Yu. I. Neimark whose good advice was given due regard in the final version, by A. N. Sklyarevich whose meticulousness is always helpful to the author, and by V. P. and F. P. Tarasenko whose suggestions made it possible to pin-point some important details in the complex object control. Valuable remarks were made by A. N. Ruban on Chapters 3 and 4, by Yu. P. Adler and A. M. Korikov on Chapter 5, and by F. M. Kirillova on Chapter 6. The author acknowledges his deep gratitude to all of them.

Foreword

... all that is simple is true,
and all that is complex is untrue.
G. Skovoroda

Each of us, at least once, was a leader, and all of us are engaged in the problems of management and its principles. Therefore, we shall state at once that we are going to discuss the control of those objects whose behaviour can be described formally using mathematics and modern computers. And let us not mix it up with management, which is the control of those objects (e.g. a state, a collective, a person, etc.) that cannot yet be formalized.

Control is something complex, delicate and vital. However, it is desirable to start the discussion on control not with its complexity, but with ... its simplicity. The author is deeply convinced that all scientific principles and concepts (including control) are always basically simple. But sometimes this kind of simplicity may turn to be deceptive.

The path from the complicated to the simple is quite clear: if something is complicated, take it to pieces, and it becomes simple and easy. But there is also another way and as a rule, it requires a formal approach, i.e. the use of mathematics. Let us not waste time making it clear why a formal approach might be needed, very often you just can't solve a problem effectively without it, and especially in the field of control.

In any case a formal approach allows us to use computers. Formalization is especially needed, in fact indispensable when you want to shift the burden of solving a problem onto a computer. This is particularly felt when solving the problem of the complex system control, it is so entangled and involved that the very idea of using a computer to 'disentangle' things seems to be saving. But this is precisely here, at the start of its formalization, that an already complex problem becomes more complicated.

For many reasons professionals (non-mathematicians) are not apt to apply this method using it only as a last resort. To begin with, only mathematicians feel at ease in

the maze of the formal approach, the professional (non-mathematician) finding it rough going since he usually takes as a departing point the essence, the physical contents of the problem, but instead he is given all kinds of equation, algorithms, transformations, etc. But that is not the end. Being at odds with all the mathematical niceties, the professional, particularly so when he is not quite sure what he wants, challenges himself, and as they say, 'mathematics triumphs over the common sense'. Unfortunately, this is not the author's imagination.

This is why, when formalizing a problem and its solution with formal methods, it is essential to keep to that saving common sense which is so valuable for informal approaches. And this is addressed to such an expert who understands the possibilities of the formal approach and has a sober view of its strong and weak points.

There is no formalization that fits everything. Our concern is the formalization necessary for control. The term 'control' is used here in the broadest sense, as a combination of actions aiming at reaching objectives which have already been set. It is easy to see that such a definition actually covers any reasonable action, but far from all of them can be formalized. This book concerns itself only with formal aspects of control that help solve the problem of the control of a complex object with the aid of computers. The social, psychological and other humanitarian aspects of control are beyond its reach, although they may change some coefficients of the equations and bring about some evident limitations.

Many efficient formal methods of control synthesis have been developed to date. They are supported both by certain mathematical models for identifications, experiment design, and mathematical programming, and by the informal methods used by experts to solve problems. Combining the formal and informal approaches helps solve the very involved problems of control and forms a basis for the control of a complex object. This combination is characteristic of the processes in controlling complex object, and therefore is the main aspect of the book. The expert method is more inherent in the first stages of control, while the formal approach is intrinsic to the subsequent stages.

Contents

From the author 5

Foreword 8

Introduction 13
I.1. What is Control? 13
I.2. Control System 18
I.3. What is a 'Complex Object of Control'? 22
I.4. Stages in Complex Object Control 25
I.5. How to Make a Cocktail? (A None-Too-Serious Illustration of the Serious Stages of Control) 36
Comments on References to Introduction 44
References 45

Chapter 1. Determination of the Objectives of Control 46
1.1. The Paradox of the Objective 46
1.2. The Model of the Subject 47
1.3. The Spaces of Situations and the Objectives 52
Comments on References to Chapter 1 58
References 59

Chapter 2. Determination of the Object of Control 60
2.1. Analysis of the Problem 60
2.2. The Method of Expert Evaluation 63
2.3. Controllability of the Object 67
2.4. The Expert Method for Evaluation of the Object's Controllability 75
Comments on References to Chapter 2 78
References 79

Chapter 3. Structural Synthesis of the Object's Model 80
3.1. Determination of the Object's Inputs and Outputs 80
3.2. Expert Ranking of the Inputs and Outputs 83
3.3. Decomposition of the Model 94

3.4. The Structure of the Model 99
3.5. Simulation Models 104
3.6. Semiotic Models 106
Comments on References to Chapter 3 112
References 112

Chapter 4. Identification of the Parameters of the Model 114
4.1. Formulation of the Problem of Identification 114
4.2. Identification of Static Objects 118
4.3. Identification of Dynamic Models 125
Comments on References to Chapter 4 129
References 129

Chapter 5. Experiment Design 131
5.1. The Basic Concepts of the Theory of Experiment Design 132
5.2. Criteria for Experiment Design 137
5.3. *D*-Optimal Design 139
5.4. Step-by-Step Design of Experiments 146
5.5. Discriminating Designs 149
Comments on References to Chapter 5 154
References 155

Chapter 6. Synthesis of Control (Decision-Making) 156
6.1. Formulation of the Problem 156
6.2. Classification of Problems of Mathematical Programming 162
6.3. Analysis of Problems of Mathematical Programming 169
6.4. Some Special Methods to Solve Problems of Mathematical Programming 173
6.5. Some Special Methods to Solve Variational Problems 182
Comments on References to Chapter 6 195
References 196

Chapter 7. Realization of Control 198
7.1. Formulation of the Problem 198
7.2. Consideration of the Effects of the Medium 199
7.3. Consideration of the Activity of the Object 201
Comments on References to Chapter 7 203
References 204

Chapter 8. Correction of the Control System 206
8.1. Correction of the Parameters of the Object's Model 207
8.2. Correction of the Model's Structure 212
8.3. Correction of the Controlled Object 215
8.4. Correction of the Objectives of Control 218
Comments on References to Chapter 8 222
References 222

Supplement
The Optimum Seeking Methods 223
S.1. Formulation of the Problem 223
S.2. Monoparametrical Local Search ($n = 1$) 225
S.3. Multiparametrical Local Search ($n > 1$) 229
S.4. Algorithms for Global Search 230
S.5. Conditional Optimization 233
S.6. Multicriterial Optimization 239
Comments on References to Supplement 244
References 245

Subject Index 247

Introduction

I.1. What is Control?

It is hardly advisable now to introduce a strict definition of the concept of control. From the time of publication of N. Wiener's book on cybernetics, this field, the theory and practice of control, has been rapidly developing, so that the very notion of control undergoes unavoidable changes due to emergence of its new facets in both theoretical and applied aspects.

American mathematician Norbert Wiener (1894-1964) is considered to be the founder of cybernetics in his own right. He called it 'the science of control and communication in the animal and the machine'. However, this definition is not full enough now. Thus, the problem of communication is already beyond cybernetics. The area of application is also wider now, it covers the society as well. There is hardly a case to put up with control in general: what is needed is in a sense the best, *optimal* control (the science of control itself serves the purpose of reaching that optimality). Therefore cybernetics is understood to date as a science of optimal control of complex dynamic systems, implying the machine, and the animal, and the society, and any other complex system. This book concerns itself mainly with technical objects of control, though its principles may be applied to control of the objects of a wider range.

We are going to understand as *control* the process of organizing such a purposeful effect on the object due to which the object is taken into the required (objective) state. This definition, just like always, substitutes one notion by others. These are the notions of 'object', 'objective', 'effect', 'state'. Although the notions are instinctively evident and understood by everybody similarly, they deserve some explanation.

Let us call the *object of control* that part of the universe the state of which we are interested in and which we can

purposefully effect on, i.e. control it. We shall designate by X the effect of the medium on the object (that is the object's input), and by Y the state of the object (that is its output). The object is shown as a rectangle in Fig. I.1.1. We shall not go, so far, into the details of these designations (that is going to be done in the subsequent chapters). It is essential only that we can speak now of the object, X and Y in natural terms, i.e. describe them in any convenient language (not necessarily formal).

For instance, if the object of control is a motor-car, then X is the state of its medium, i.e. roadway pavement, air, lighting, as well as the position of its control organs (the driver is also a part of the medium for the car). Y is the state of the car, including essentially its speed, oil pressure, temperature and rpm of the engine, etc. It is obvious that the state of the object Y depends on the state of the medium X, i.e. there is some relationship between them, the nature and form of which are not known yet. This is represented diagrammatically in Fig. I.1.1, and it may be shown as a formula:

$$Y = F^o(X), \tag{I.1.1}$$

where F^o means the relationship between the object's input and output. In mathematics such a relationship is called a function for static objects, and an operator for dynamic ones. Therefore, in the most general case, F^o is, so far, an unknown operator transforming the input X into the output Y of the object. This operator characterizes specificity of the object from the standpoint of control.

Speaking of control as a purposeful process, one cannot avoid considering whose objectives are reached in the process of control, i.e. who needs that control. Therefore, it is quite natural to introduce into discussion the *subject* who is the source of the objectives to be reached by control. The subject's objectives emerge due to his own needs which are related to his living activity and interaction with the medium and the object of control. (Let us note that a subject is not necessarily a person; that may be a group of persons with the same interests, and even the whole mankind, in case the object of control is global, such as environment, outer space, etc.)

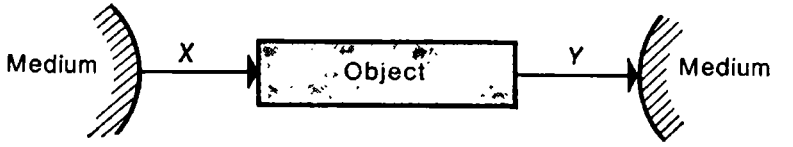

Fig. I.1.1. Interaction between the object and the medium

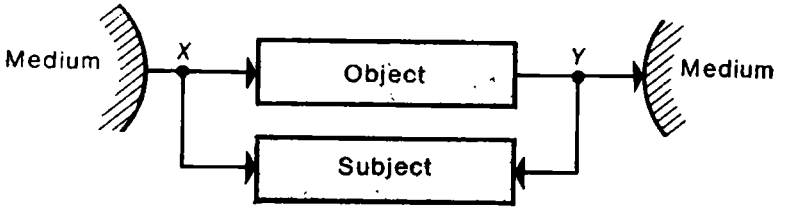

Fig. I.1.2. Interaction of a subject with the medium and the object before control

In the course of the analysis of control, the emergence of the subject with his needs is absolutely necessary, if we want to understand how the objectives of control are formed.

Speaking of the purpose of control, one should bear in mind the well-known rule of detectives investigating a crime: 'Search for who profits by it!' Applying this rule to control, we come ... to our subject: he is the only figure who needs control and profits by it.

Since the figure of the subject is so essential for control, let us isolate him from the medium, as it is shown in Fig. I.1.2. The subject is in the same medium as the object, i.e. he is capable of sensing the state X of the medium. At the same time, and that is the main point, the state Y of the object effects him as well. In case the state Y of the object meets the requirements of the subject interacting with the object and applying it for his ends, no control is needed. But if the subject cannot put up with that state, he has to produce such an effect on the object which is to take it into a new state, satisfactory for the subject. This effect is an act of control. Hence, one of the 'principles' of control: it results from the subject's dissatisfaction with the current state of the object.

If the subject 'prospers', i.e. the object satisfies all the needs of the subject, there is no need in control. But it is convenient (however conventional that may be) to assume

that the subject always states his objective so that its implementation in the object would result, in the opinion of the subject, in satisfaction of all his needs. This objective is a set of the subject's requirements on the state of the object.

Let us designate the objective of the subject by Z^* (all that is desired will be symbolized further with an asterisk). Basically, this objective determines the way the object should be from the viewpoint of the subject. To check whether the objective Z^* is reached in the object, it is necessary to know its state Y, but first it is necessary to express this state Y in the language of the subject's objectives, i.e. to carry out the transformation

$$Z = \Psi(Y). \qquad \text{(I.1.2)}$$

It may turn out in a particular case that $Z = Y$, i.e. the subject puts forward his objectives in the language of the states of the object. For instance, let the objective Z^* be to maintain the speed of the car 60 mph. The speedometer in the dashboard reads the actual speed Y in the same units (mph). This is the case when $\Psi(Y) = Y$, since there is no need to 'translate' the readings of the speedometer into the language of the objective: that has been done already. But if you drive another car, where the speed is shown in kph, then $\Psi(Y) \neq Y$. Now $Z = 1.61\,Y$, i.e. there is a need in translating the registered state of the object Y [kph] into the language of the subject [mph].

Evidently, the equality

$$Z = Z^* \qquad \text{(I.1.3)}$$

shows that the state of the object satisfies the objectives of the subject, i.e. the objectives of the subject have been attained. But if

$$Z \neq Z^*, \qquad \text{(I.1.4)}$$

then the objectives of the subject are not reached in the object. That compels the subject to solve the dilemma: he has either to put up with Z and tolerate a certain loss due to his inability to reach the objective Z^*, therefore leaving some requirements unsatisfied, or to create a system of control that would be able to realize his objectives Z^*

in the object, though spending certain resources to make and to run it. (Here a 'system of control', or 'control system', implies all the measures necessary to produce and realize the control.)

We can tolerate this in the majority of situations. The principle of patience is proverbial. Strictly speaking, patience is also control, but inward, directed to the subject's self-restraint and limitation of his needs. That is the control over the requirements. To a certain extent that is reasonable. As a matter of fact, if one can have patience, there is no urgent need in control. So control is the last resort.

The trouble is that one has to pay for control and not uncommonly an exorbitant price! That concerns expenses for collecting data on the object, synthesis of its model, producing control measures and their realization, etc., etc., etc. That is precisely why control is the last resort. However, such extreme circumstances occur far too often (ever increasing demands of the subject are here to blame). And that brings us to the development of general methods of control reviewed in this book.

In case the objectives are essential (i.e. demands which engendered them are really urgent), the subject takes pains to create a system of control. First, he has to define the ways to influence the object, i.e. define the channels of control. These may be some of the inputs X of the object lending themselves to purposeful change. But not uncommonly that does not suffice, one has to establish new control channels, non-existent before. Thus, to be able to change the course of the cart, man had to develop the steering wheel, the means to alter the orientation of the wheels the cart travels on. And that was a new control channel.

Somehow or other, special control channels are established to realize control. Let us designate that control by the letter U. It is obvious now, that the state of the object depends on two factors: the state of the medium X and the state of control U:

$$Y = F^0(X, U), \tag{I.1.5}$$

where F^0 is the same operator of the object, now taking into account also the control U.

This is the point where we can speak of creating a *control system*, by which we shall imply all the pertinent algorithms for data processing and the tools for their realization combined to reach the given objectives of the object control. It should be noted that a control system is far from always realized as hardware. It may be realized as a system of rules, agreements and liabilities (i.e., generally, algorithms) carried out in the process of control. Thus, the control system consists of an operator together with the object which the operator controls. For instance, a car-driver at the steering wheel together with the car itself is a control system. In this case the means to control the car (the steering wheel, foot throttle, brakes, etc.) are the very channels of control U which allow the car-driver to change the state of the car Y so that to reach the desired goal Z^*. This is when the 'control device' (the car-driver) and the subject setting up the objectives Z^* are the same.

I.2. Control System

A diagram of the control system is shown in Fig. I.2.1. Here S_X and S_Y are the sensors to read the state of the medium and the object respectively. The readings

$$X_d = D_X(X); \quad Y_d = D_Y(Y), \tag{I.2.1}$$

where D_X and D_Y are the operators of the sensors, are the original data for the control device (CD) to produce the control U.

The formulas present the evident connection between the state (of the medium and the object) proper and the data on the state. And those are a far cry from each other. The difference is easy to see if you try to study the state of the Black Sea by the data in the photos and stories about it. Now X is the actual state of the Black Sea, and X_d are the photos and stories.

We can feel all that surrounds us with our five organs of sense, and we act (realize control) on the basis of these sensations. If we want to make a computer responsible for control, it should be supplied with 'organs of sense', a set of the sensors S_X and S_Y to read the state of the medium X and the object Y.

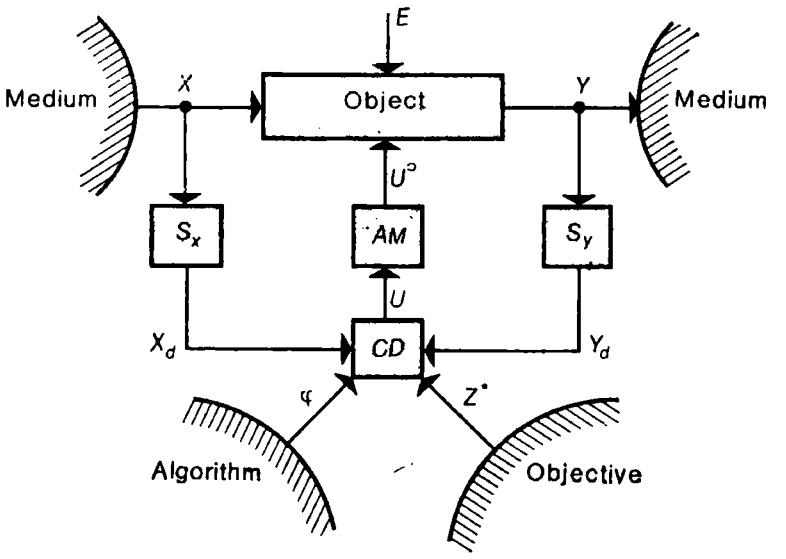

Fig. I.2.1. Diagram of the control system

Consequently, the process of control is underlied by the information on the situation

$$I = \langle X_d,\ Y_d \rangle, \tag{I.2.2}$$

although it is always incomplete. This incompleteness is primarily due to limited possibilities of any system for collecting the information I and due to the need to 'pay' for the data. Considering limited resources allocated commonly for control, this brings about a standing deficit of the data on the behaviour of the medium and the object. Since the information on X and Y is accessible in the form of X_d and Y_d only, it is natural for simplicity's sake to assume $X_d = X$ and $Y_d = Y$ which is going to be meant further.

Consequently, we are compelled to take into account the existence of still another input to the object, the unobserved disturbance E (see Fig. I.2.1), which implies all the unobserved external and internal, with regard to the object, factors effecting its state Y. So $Y = F^o(X, U, E)$, which makes (I.1.5) still more accurate.

One may only have certain assumptions with regard to the input E, but there are no direct measurements. The

input E includes unmeasured parameters of the medium (e.g. the temperature of the environment in automatic control of the temperature inside the refrigerator chamber) and various random changes in characteristics of the object (e.g. their drift).

Let us note that the control actions U (instructions produced by the control device CD) are the data on how to set the controlled inputs U^o of the object. 'Execution' of the commands is carried out by the actuating mechanisms AM that change the state of the controlled input U^o of the object.

Therefore, the object of control has three inputs: the observed input X, the controlled and therefore observed U^o and the unobserved E. To simplify the matter, let us assume that the actuating mechanisms carry out the instructions U instantaneously, and therefore $U^o = U$, i.e. there is no need in making distinctions between U^o and U (the case when $U^o \neq U$ is analyzed in the theory of automatic control, to be more precise, in the theory of servo systems, well developed and covering the field substantially).

To ensure its purposeful functioning, the control device has to 'know' apart from the data (I.2.2) the objective Z^* of control, i.e. what it should aim at, and the algorithm of control, i.e. an instruction how to reach the objective.

Thus, very essential concepts have been introduced: those of the *objective* and the *algorithm*. While the notion of the objective is intuitively clear (that is what we are striving for), that of the algorithm should be explained. An algorithm in this case means a distinct unambiguous set of rules, instructions, directions as to what should be done and in what way so that to achieve the set objective Z^* in any of the situations I.

For instance, any instruction attached to a drug is an algorithm, since it specifies the cases (I) when the drug should be applied and when not. Here the objective (Z^*) is recovery of the patient, and the control is the dosage of the drug (zero dosage is also a control). However, an instruction 'go somewhere else and fetch something or other' cannot be said to be an algorithm, though it has a similar structure.

When the data on the objective and the state of the medium and the object are available, it is possible to present the control U as the realization of the algorithm

$$U = \varphi(I, Z^*), \qquad \text{(I.2.3)}$$

where φ is the algorithm of control. It is an operator $\varphi = \varphi(\cdot, \cdot)$ to process the data on the medium, object and objective into the control U. The latter, realized as U^0, should take the object into the required state Z^*.

Consequently, the basic factors of any control, inherent in any system to control an object, are:

— the objective of control (Z^*);
— the information on the state of the object and medium (I);
— the action exerted on the object, i.e. the control proper (U);
— the algorithm of control (φ).

Let us note that the term 'control' is commonly used in two senses. In its widest sense it is a process to reach some pre-set goals (this is the sense we use), and in the narrower sense it is any action exerted on an object (pressing a button, turning a wheel, allocating resources, etc.).

If you except from the list of the basic factors at least one of them, control will become impossible. During hostilities, for instance, this is precisely the tactics employed by the subversion units striking the system of control over the enemy troops. Regular objects of diversion are the communication systems and the headquarters: the communication system is the means for conveying the information I and the control actions U, while the headquarters realize the algorithm φ of control over the troops.

Thus, control in the wide sense is defined by a set of objectives $\{Z^*\}$ coming into the control system from without ($\{a\}$ taken in braces will denote the set of the elements a). The objectives are put forward by the subject, who is going to be the 'consumer' of the control system. He is the client ordering the development of the control system. And the development of the system is provided by a designer.

However, before taking the decision to develop a control system, it is necessary to consider all the stages of

control irrespective of what material means are going to be applied to realize those stages. Such an algorithmical analysis of control is the basis for making the decision on the development of the control system and the degree of its automation.

Such an analysis should regard the factor of 'complexity' of the object of control. As a matter of fact, control over a 'simple' object is no problem nowadays. An example of control over a simple object is the process of automatic stabilizing the temperature at a given (objective) level. This process is realized in any refrigerator. The algorithm to control the temperature is very simple here: switch on cooling if the temperature inside is more than a given one, and switch it off if less or equal. And that's all!

Every problem of control springs from the complexity of the object, and therefore one should start the control with the object, and not with the algorithm. Let us stop to think which object of control shoud be called *simple*, and which *complex*—without reference to control itself. Unfortunately, a short definition just won't do here. Therefore, we shall have a whole section to discuss it.

I.3. What is a 'Complex Object of Control'?

This concept has not been strictly defined so far. Possibly, such a definition will never appear and if it will, the chances are there will be no urgency in it then. The intuitive notion of a complex object rather closely correlates to the concept used in the theory of control (and that is what is going to be used in this book). Let us consider some features (but not the formal attributes) of a complex object. They are many. Here are the basic ones.

1. The absence of a mathematical description and the need in it. That is an inherent feature of a complex object of control. A mathematical description implies an algorithm (a rule, an instruction) for computing the state Y of the object according to observations of its inputs, both controlled U and uncontrolled though observed X. From this point of view a modern ballistic (or space) missile is not a complex system, because its movement

may be described by the equations of classical mechanics (Newton's laws), which are easy to set up and solve, since the parameters of the missile and the properties of the medium are known. Therefore, the behaviour of the missile is more definite than that of an ant which should be considered to be a complex system by all means, since no one even tried to delineate the ant's behaviour in the terms of mathematics.

Simple objects of control may also have no mathematical description. Thus, to control the temperature in a room, there is no need to know the mathematical relationship between the temperature of the heater (cooler) and the temperature in the room. However, while simple objects may be controlled without their mathematical model, complex ones may not. Thus, when there is a need to maintain a given temperature in a room precisely, you cannot do without a model. The requirement of accuracy renders a simple object a complex one. This inference leads to the following: if there is a need to control a complex object using some formal methods, you have to develop its mathematical model. This is a natural cooler to the zeal of some hotheads to control everything on earth with the help of the computer.

2. Stochastic behaviour of the complex objects of control. That is also an essential feature, characterizing the problems of their analysis and control. This feature is due not so much to some special sources of noise within the object as to the very complexity of the object and inevitable abundance of all kinds of secondary (from the standpoint of control, of course) processes related to it. Therefore, its behaviour is often unexpected to the researcher, and this unexpectancy is better to regard as internal noise than go deep into the mechanisms of the minor processes in the complex object. Any complex object presents a host of such unexpected things which are only the evidence of its complexity. Such are biological, social, technological, and many other objects of research and control.

This is a display of a peculiar tendency to call incidental what is minor or inessential for the realization of the objectives of control in the object. On the contrary, those 'incidental' things may turn to be not incidental at all.

But ours is the case when it is very convenient to consider incidental all that is unclear. This kind of approach is well developed in the statistical theory of control (let us note that the terms 'statistical' and 'stochastic' are tantamount).

3. 'Intolerance' to control. That may be the most vexing feature of the complex object impeding its control. As a matter of fact, a complex object exists and functions independently from the subject and his demands. Thus, any technological process is meant for turning out products, not for control over it. Control is something external with regard to the object. It is natural then that due to this any control disturbs the 'normal' activity of the object, i.e. introduces some changes into its independent behaviour and makes it dependent on the subject. This is especially obvious in active objects with human beings or collectives in them (e.g. when the object of control is a technological line, a workshop or a factory). It is difficult to count here on that the own objectives of such a complex object coincide with the objectives of control. Rather, they are going to be at some variance to each other. And that is what brings forth an adverse reaction of the complex object, if the objectives of control are not matched with its own objectives.

4. Non-stationarity. That feature of the complex object results from its very complexity. It reveals itself in the drift of the characteristics of the object, in their 'waving', i. e. in the evolution of the object in time. The more complex the object, the more salient is this feature seriously curbing the synthesis of the model of that complex object and control over it. The more complex the object, the faster it changes. In our day-to-day life, we observe it in the depreciation (ageing) of the things around us. And as a rule, the depreciation of complex things goes on more rapidly than that of simple ones. Thus, a TV set or a car 'grow older' quicker than a hammer.

The following feature of the complex object is a consequence of its non-stationarity.

5. Irreproducibility of experiments. This feature manifests itself in different response of the object to the same circumstances or control in different moments of time. All the time the complex object denies itself, as it were,

i.e. changes constantly. This must be taken due regard of in the synthesis of the model of the object of control, for which a special correction is introduced.

All that results in that the objective of control can never be reached in full. As a matter of fact, some time is necessary to synthesize the control U, during which the object is going to change in a way unpredictable to a certain extent, so the control in all probability won't bring about the desired Z^*.

An effective means to counter the listed properties of the complex object of control is the extrapolation of the system's behaviour, i.e. projection of its evolution. Now the control U may be carried out with a lead, anticipating some change in the object. Another method is to reduce the cycle of control, so that the object would not change much during the period for the control synthesis.

The list of the 'vexing' features of the complex object of control is far from exhausted. However, it should be borne in mind that those are the features inherent in the complex object, but not its formal properties by all means. The absence of one or even several of the above-mentioned features does not necessarily render the object simple. Those are the informal features characterizing any complex object of control.

I.4. Stages in Complex Object Control

Control over the complex object goes through the following stages (Fig. I.4.1).

1. *Determination of the objectives* $\{Z^*\}$ *of control.* This is the stage when the objectives (more accurately the set of the objectives) to be realized in the process of control are determined. The notion of 'objective' is used here in the sense of the 'model of the future', i.e. some presupposed state of the object, which is desirable to the subject (consumer), and which is in a sense 'unnatural', i.e. cannot be realized in a natural way in the medium without external interference, i.e. without control. Simply speaking, it is at this stage that it is cleared out what is wanted by the subject and whether it is worthwhile to want. And in case it is worthwhile, whether the desire may or

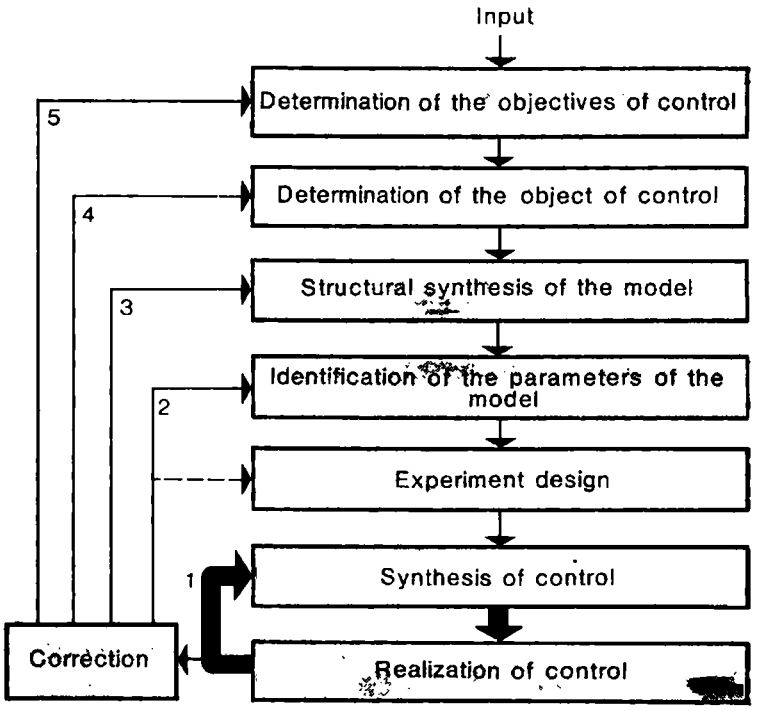

Fig. I.4.1. Stages in the control over the complex object

may not be realized by itself. And if not, the desire becomes an objective of control.

Thus, maintaining the temperature t in a room at a certain required level t^* (whatever the weather) is impossible without control. The control may be automatic (by an air conditioner) or non-automatic (manual), but in either case the objective of such a control is Z^*: $t = t^* = \text{const}$, while the constant t^* may change depending on the circumstances (the needs of the subject).

Setting oneself a number of objectives $\{Z^*\}$ in the given case reduces to the determination of the range of temperatures, which should contain the objective (i.e. necessary for the subject) temperature t^*, whenever and whatever the unknown circumstances might happen to be. Automatic control in this case should be a hardware realization of the following algorithm: to switch over to the heater when

$t < t^*$, and switch over to the cooler when $t > t^*$. This is the way the air conditioner should work. But the same algorithm may be realized manually, without automatics.

2. *Determination of the object of control.* This stage is the isolation of a certain part of the medium of the consumer, the state of which part he is interested in to realize the predetermined objectives. That part is the object of control. For instance, the determined range of the objective temperature implies an object of control, i.e. the part of the universe the state of which should be checked (the temperature measured) and which should be acted on (heated or cooled in places) to achieve the determined objectives of control $\{Z^*\}$.

The object of control is often clearly determined from the start (a car, a radio set, a factory, etc.). But when there are numerous and essential connections with the medium, the object of control may not be easy to determine. Try and determine the sphere of your control (i.e. the object) to keep your apartment clean. The apartment? But the dirty staircase makes you lodge a complaint. And that is also a control the objective of which is the same: to keep your place clean. So the object should be expanded and include the staircase and the house management. The absence of asphalt paths near the house makes you appeal to the municipal authorities, and that aims at the same: to keep your home clean, i.e. your object of control includes then your house yard, the related department of the district authorities, etc.

The next three stages of control are linked to the development of the model of the object of control, which is necessary for the synthesis of control.

Let us imply by the model of the object a dependence F of the state Y of the object on its inputs, both uncontrolled X and controlled U:

$$Y = F(X, U). \tag{I.4.1}$$

Unlike the unknown operator F^o of the object in (I.1.5), the model F is a statement regarding the connection between the inputs (X and U) of the model and its output in any convenient language. Generally, the model F is defined as an algorithm (a rule, an instruction) which indi-

cates how, having the data on the inputs X and U, to define the output Y without the real object.

It is only with the aid of the model of the object that it is possible to produce the control U in order to take the object into the required (objective) state Z^*. Without the model the process is most likely to be realized by the trial-and-error method which is unacceptable in the control over the complex object, since it requires too much time and makes one 'jerk' the object by wrong actions (steps).

The process of the synthesis of the model commonly falls into three stages.

3. *The structural synthesis of the model.* This is the stage when the form of the dependence F is defined with no regard to the values of its parameters.

Let us carry out a theoretical operation: 'split' the model F into its structure St and parameters $c_1, \ldots, c_k$, i.e. present the model as a pair

$$F = \langle St, \mathbf{C} \rangle, \qquad \text{(I.4.2)}$$

where $\mathbf{C} = (c_1, \ldots, c_k)$ is the vector of the parameters of the model (we shall give vectors in the medium-boldfaced type). At the stage of structural synthesis only the structure St of the model is defined, with no regard to concrete values of the parameters $\mathbf{C}$.

Usually the structure is understood as the kind of elements of which the object consists, and the kind of relationships between the elements. It is most convenient to show the structure as a graph, the nodes of which are the elements of the object and the branches, their interrelationships. Thus, a factory as the object of control includes workshops, departments, services, etc. (those are its elements), which are interrelated by subordination, material and information flows, etc.

The structure of the object of control may be different depending on the objectives of control. Let us illustrate this point with the following (not very much serious) example. One and the same object, a factory, may have different structures from various points of view. Thus, from the point of view of the related ministry, the factory is a converter of the allocated funds (both financial and material) into the output products. The workers of the fac-

tory consider it as a converter of their labour into wages (that is a deliberate simplification). Medical institutions regard the same factory as a source of various occupational traumas and diseases, and the sanitary department is convinced that the factory produces environment polluting wastes.

There may be a variety of structures of the models of the objects of control. Linearity, static nature, determinancy, discreteness, etc. are structural categories. For instance, any linear static continuous determined structure unambiguously defines the following form for F:

$$Y = F(X, U) = c_0 + c_1X + c_2U,$$

where the concrete values of the parameters c_0, c_1, c_2 are not essential so far, but the form of the dependence F is essential, meaning linear nature of the structure St.

Consequently, at the stage of the structural synthesis, only the general nature of the model F is defined, while its parameters C are defined at the next two stages of control.

4. *Identification of the parameters of the model of the object.* This is the determination of the numerical values of the parameters $\mathbf{C} = (c_1, \ldots, c_k)$ in the mode of normal functioning of the object, i.e. without special control actions*. The initial information for identification is the structure St and observations of the input $X(t)$ and output $Y(t)$ of the object in its interaction with the medium. Therefore, the pair

$$I(t) = \langle X(t), Y(t) \rangle \qquad \text{(I.4.3)}$$

is generally the main data source in identification.

The art of identification is the art of passive observation. That was the basic credo of the ancient science: to be able to observe without interference. This is supported by the fact that any interference distorts natural evolution of the observed phenomenon. Moreover, the results of an experiment are somehow predetermined by its goals. Very

* 'Identification' sometimes implies the process of determination of the parameters of the model irrespective of its model of activity. Therefore, passive and active identification are distinguished. Only passive identification is considered here, an experiment design being active identification (see the next stage).

often the experiment yields not what there is, but, roughly speaking, what is very much desired, and that is not because the experimenter is unscrupulous, but due to the psychology of the process of cognition. That is why passive observation still stays an effective instrument to cognize the surrounding world.

The need in passive observation is also related to the fact that complex objects often do not allow experimenting with them. This is apparently what produced the absurd notion of 'passive experiment'. An experiment cannot be passive, it is always active! The idea is just passive observation.

As it is easy to see, information (I.4.3) allows the determination of the state of the object depending on the input X only: $Y = F(X)$, but such a dependence differs from the sought model (I.4.1). To find out the dependence of the output Y of the object on the controlled input U, it is necessary to change it intentionally, i.e. to experiment with the object. However, as it was mentioned above, the complex object 'does not like' experiments interfering with its normal activity. Therefore, any unavoidable experiment should be carried out with minimal disturbance to the object, aiming nevertheless at the maximal amount of the information on the influence of the variable parameters on the output of the object. That is what the next stage is for—experiment design.

5. *Experiment design.* This is the stage where the point is the synthesis of the experiment design allowing the determination of the model of the object of control with maximal efficiency.

This design (plan) $\tilde{U}$ for the static object is a set of the states of the controlled input of the object

$$\tilde{U} = \langle U_1, \ldots, U_N \rangle, \qquad \text{(I.4.4)}$$

and a design-function for the dynamic object

$$\tilde{U} = U(t), \quad 0 \leqslant t \leqslant T, \qquad \text{(I.4.5)}$$

i.e. a program of changing the controlled input of the object in time. An experiment with the object helps define the reaction of the object to that control action. For the

static object, this reaction forms a sequence

$$Y_1, \ldots, Y_N, \tag{I.4.6}$$

where $Y_i = F^o(U_i)$, $i = 1, \ldots, N$, is the reaction of the object in the i-th experiment, and for the dynamic object it is $Y(t) = F^o(U(t))$. Now N pairs I_i:

$$I_i = \langle U_i, Y_i \rangle, \quad i = 1, \ldots, N, \tag{I.4.7}$$

for the static object, and a pair I:

$$I = \langle U(t), Y(t) \rangle \tag{I.4.8}$$

for the dynamic object are the initial information for the determination of the parameters C of the model F:

$$Y = F(U). \tag{I.4.9}$$

Obviously, the problem of determining the parameters of the model after the experiment is reduced to the problem of identification (see the preceding stage).

To obtain a model of the (I.4.1) type, it is necessary now to consider the 'threes' I_i of observation:

$$I_i = \langle X_i, U_i, Y_i \rangle, \quad i = 1, \ldots, N, \tag{I.4.10}$$

or

$$I = \langle X(t), U(t), Y(t) \rangle, \quad 0 \leqslant t \leqslant T, \tag{I.4.11}$$

i.e. it is necessary to take into account the state X of the medium when designing experiments with the object.

Notably, during the process of designing experiments with the object, the experimenter may change his concept of the structure of the model (e.g. an initial hypothesis of linearity of the model may be replaced by a hypothesis of non-linearity of a certain kind). This circumstance compels to turn back to the structural synthesis, or more precisely to introduce correction of the structure of the model. The said somewhat 'diffuses' the concept of experiment design and spreads it over the processes of both making out the structure of the model and its correction.

Experiments may be designed not only for pin-pointing the parameters of the model, which has already been discussed, but also for selecting a structure out of a number of presupposed ones, etc. In case, for instance, the object is a person, it is possible to design experiments to find out some of the person's specific qualities which are of

interest for control (e.g. in looking for occupational tendencies). Such experiments are psychological tests. The tested person is given a number of problems (tests) to solve, and the answers help in defining his (or her) inclinations, i.e. a model of the person being tested is shaped. Such tests are compiled with the aid of the methods of experiment design, which are to help find out the 'belonging' of the tested persons to a certain vocational structure.

Consequently, at the stage of experiments design, those parameters of the model are determined which cannot be determined at the identification stage, and the structure of the model may be corrected. However, the aim of both stages is to produce a final version of the model F of the object F^o of control.

The problem of the model synthesis which was tackled with in the last three stages (3, 4 and 5) may be considered to be solved. The produced model is an initial point for the synthesis of control.

6. *Synthesis of control.* This stage includes making a decision with regard to what kind of control should be U so that to reach the desired objective Z^* of the object control. The decision should rest on the available model F of the object, the given objective Z^*, the acquired data on the state X of the medium, and the allocated resource R of the control, the resource most commonly being a set of limitations imposed on the control U due to specific properties of the object and the possibilities of the control system (concerning time, energy, material factors, etc.).

It is natural to suppose the synthesized control U to be optimal, i.e. $U = U^*$, since it should satisfy the desired objective Z^* of control. Generally, it is a programme of the change of the controlled parameters in time, i.e.

$$U^* = U^*(t). \qquad \text{(I.4.12)}$$

This programme is to be realized at the next stage.

An example of a programme of optimal control which everyone is well acquainted with is a plan for medical treatment prescribed by a physician: 'Well, that's what we'll have...', and then a detailed plan U^* of the treatment considering: the objective Z^* (to render the patient 'practically' well with minimal damage due to the treat-

ment, i.e. minimal risk of aggravation, etc.), the state X of the medium (the season of the year, patient's living conditions, whether there are relatives to look after the patient, etc.), the resource R of the patient (a possibility to get some deficit drugs) and that of the clinic (availability of the necessary instruments and equipment for making diagnosis and treatment), the model F of the object, i.e. the physician's concept of the patient at the given moment from the medical viewpoint (i.e. the diagnosis, the results of the tests, tolerance of the patient to certain drugs, whether he is disciplined enough, his psychic condition, etc.).

7. *Realization of control.* This is the realization of the programme of the optimal control (I.4.12) produced at the preceding stage. The process is no problem if the data on the state X of the medium, Y of the object, and F of its model be reliable. However, all those may change by the moment of realization. To take due regard of such changes, it is necessary to make a correction of the control U in the process of realization. Initial information for such a correction is the data on the new state of the medium by the moment of realization of control, or a forecast of that state. If the object is active, e. g. contains human beings whose objectives may not be the same as the objectives of control, this should also be considered in the process of control realization. Apparently, this is a case when the desired objectives cannot be attained in full, so the actions should render the discrepancy minimal.

That is why the design for optimal control $U^*(t)$ and its implementation often differ just as our good wishes and their realization. Implementation of control is often more troublesome than its synthesis. For instance, designs of optimal diets for slimming are always discarded precisely because they require implementation, i.e. limitations unbearable for fat people. Disregard of that point renders the control unrealizable.

After realizing the control and acquiring a new state of the object, it is possible to find out whether the desired objective has been reached. It is hard to expect that it has, since the object is complex, non-stationary and with internal noise, due to which the one-act control cannot

bring about the goal. Therefore it is neccesary to return to one of the preceding stages. Even at its best, when the desired objective has been reached, there is still a necessity of turning back to a preceding stage because of a change in the state X of the medium or a change of the objective Z^* of the control. Consequently, even the most favourable circumstances compel us to return to the stage of the control synthesis to determine a new control reflecting new conditions of the medium and a new objective. This is the way any normal control loop works (see arrow *1* in Fig. I.4.1).

8. *Correction.* This stage, due to the very complexity of the object, is a return to one of the preceding stages of control (see arrows *2-5* in Fig. I.4.1). The trouble is that all the decisions taken at those preceding stages are based on outdated information and reflect the state of the object in the past.

Correction may cover various stages. The most common correction is the resetting of the parameters C of the model (see arrow *2* in Fig. I.4.1). This kind of correction is usually called the adaptation of the model. It may turn out sometimes that the optimal control $U^*(t)$ does not provide a necessary variety of the controlled input of the object for subsequent effective correction of the parameters C of the model. Then it is necessary to turn back to the experiment design right in the process of control by means of adding special test signals to the control (see the dashed arrow in Fig. I.4.1). Such a control is called dual, or double, since its aim is not only to reach the objectives of control, but also correct the parameters of the object model.

However, the correction of the parameters C of the model may prove to be not enough in case the structure of the object has changed. Therefore, a correction of the structure St of the model may be needed from time to time, which means bringing it into line with the changing structure of the object (see arrow *3* in Fig. I.4.1).

Furthermore, correction may be pertinent to the object itself, to be more precise, to the boundary between the object and the medium. That is quite common in a considerable change (evolution) of the object and the medium. Correction in this case is going to 'legalize' the new rela-

tionship between the new object and the medium (see arrow *4* in Fig. I.4.1).

And one more point. There may be a number of reasons, such as an essential change in the properties of the medium and the object, why the produced system of control will not be able to realize the whole of the initial set of the objectives $\{Z^*\}$ of control that engendered the system or it will be able to realize them insufficiently. Then, there is a need to correct the objectives of control, i.e. to determine a new set of the objectives to be reached by means of that system of control. That kind of correction enables the consumer to learn the objectives the control system can realize in the object effectively (see arrow *5* in Fig. I.4.1).

The realization of the above-mentioned measures sets up an adaptive control system that can adjust to the ever changing properties of the medium, object and to current needs of its consumer. Curiously, the consumer is quite commonly satisfied with the obtained control system despite the fact that it does not realize all his initial objectives. An opportunity to attain some extra objectives (previously unconsidered) changes his needs with respect to those easily achievable objectives. True, we often have a meal not because we are hungry, but because of the circumstances, and 'the appetite grows by eating'. The ease of achieving the nourishment goals itself gives rise to these goals. This mechanism is especially evident in those cases, when widespread advertising of supply results in demand, i.e. the need.

Due to the system of correction, the control system is being constantly brought into line with the ever changing medium and the object of control. It follows that the control system evolves, as it were, along with the object and the medium, so that the objectives of control should always be realized well. If the evolution of the object advances so far that the initial set of the objectives cannot be reached any more, then there occurs correction of the objectives of control. Let us note that the more the number of the arrow in Fig. I.4.1, the less often the feedback operates.

Obviously, not always all of the above-mentioned eight stages of control are included into the synthesis of the

control system. In a number of cases some of them may be absent, for instance, when the object of control has already been separated from the medium (an aircraft, a spaceship, etc.), or when there is no need in the stage of experiment design, since the model of the object is so simple and 'apparent' that all of its parameters may be defined without a special experiment. The stage of determination of the objectives of control may fall out as well, if those are quite evident and natural. Nature is not very much 'insidious', and real control systems rarely have to be worked through all the stages of control. As a rule, some of the stages fall out by virtue of a peculiar character of the objectives and the object of control. Thus, there is no problem in the determination of the object of control while driving a car, and when a patient is treated the objective is always the same, to cure and not to harm. However, the stage of correction of the model of the object should always be present in any complex object control. That stage is called for by the very complexity of the object of control. It may be avoided only if the object is 'simple'. However, let us be ready for the worst and discuss all the eight stages locked in the 'iron ring' of the control system!

I.5. How to Make a Cocktail? (A None-Too-Serious Illustration of the Serious Stages of Control)

Let us illustrate the above-mentioned stages of the complex object control by a joking problem of cocktail-making.

Suppose you are facing a problem of making a cocktail, nice to taste and strong enough on the one hand, and not too expensive on the other. You will have to solve the problem on what and how much you should buy to reach the target. But what is that target? What specific objectives should be reached?

The objective criteria of any cocktail are commonly its taste, strength and cost per unit volume. Suppose z_1 is the taste of the cocktail, z_2 is its strength (e. g. in per-

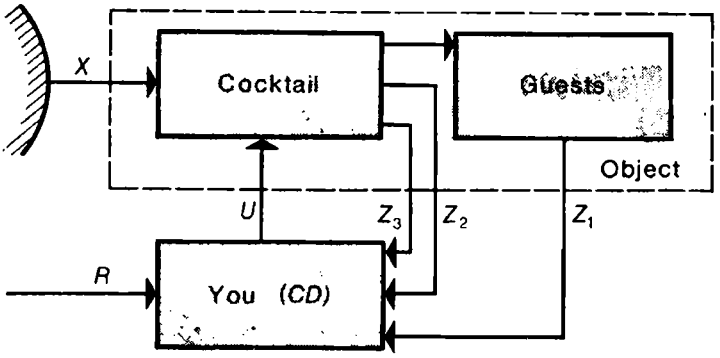

Fig. I.5.1. Diagram of a 'control system' for cocktail-making

centage of alcohol content), z_3 is the cost of a litre of the mixture (e.g. in roubles), u_i is the relative content of i-th ingredient in the cocktail, where, for instance, $i = 1$ means fruit juice, $i = 2$ is dry wine, $i = 3$ is cognac, etc. All in all, there are n ingredients ($i = 1, \ldots, n$), so that $\sum_{i=1}^{n} u_i = 1$; while $u_i \geqslant 0$.

The vector $\mathbf{U} = (u_1, \ldots, u_n)$ represents the contents of the cocktail. All the three criteria definitely depend on the vector, but the nature of the dependencies is not known yet:

$$z_1 = f_1(\mathbf{U}), \quad z_2 = f_2(\mathbf{U}), \quad z_3 = f_3(\mathbf{U}),$$

where f_1, f_2 and f_3 are the so far unknown objective functions.

Let us delineate the problem of cocktail-making as a problem of control. The object of control in this case is the cocktail and the guests who provide the appraisal (see Fig. I.5.1), the uncontrolled input **X** is the basic ingredients of the cocktail, the controlled input **U** is their relative contents. The output of the object is the properties of the cocktail $\mathbf{Y} = (z_1, z_2, z_3)$, your guests being the savourers, and the 'control device' (CD) is the bartender (that means You).

Let us first determine the objectives of control (stage 1). It has already been said that they depend on the demands of the consumers of your cocktail, your guests.

If the guetss have a sweet tooth, they concentrate on the taste of the cocktail, so in this case it is natural to

consider the criterion z_1 (the taste) to be extremal which (to be more precise, the objective function f_1 (U)) should be maximized in the process of control. The criterion z_2 (the strength) should be bounded below: $z_2 \geqslant a_2$, where a_2 is the minimal strength to suit your guests. The criterion z_3 (the cost) should be bounded above: $z_3 \leqslant a_3$, where a_3 is the maximal cost of a litre of cocktail specified by your budget. So the objective with regard to gourmets may be defined as

$$\mathbf{Z}^*: \{z_1 \rightarrow \max;\ z_2 \geqslant a_2;\ z_3 \leqslant a_3\}.$$

That means that the gourmets will be satisfied if the taste of your cocktail is maximized while its strength and cost are bounded with a_2 and a_3.

The target aimed at boozers is

$$\mathbf{Z}^*: \{z_1 \geqslant a_1;\ z_2 \rightarrow \max;\ z_3 \leqslant a_3\},$$

i.e. to maximize the strength of your cocktail bounding its minimal tolerable taste with a_1 and its maximal cost per litre with a_3.

The objective of a hard-up host trying to minimize the costs may be defined similarly as

$$\mathbf{Z}^*: \{z_1 \geqslant a_1;\ z_2 \geqslant a_2;\ z_3 \rightarrow \min\}.$$

Evidently, to determine any of the objectives, it is necessary to set two of the three figures: a_1, a_2, a_3.

Consequently, the problem for gourmets may be presented as follows:

$$f_1(\mathbf{U}) \rightarrow \max_{\mathbf{U} \in \Omega_1},\quad \Omega_1: \begin{cases} f_2(\mathbf{U}) \geqslant a_2, \\ f_3(\mathbf{U}) \leqslant a_3, \end{cases} \tag{I.5.1}$$

i.e. you should ($\rightarrow$) maximize (max) the function f_1 (U) for the taste, with certain constraints on the functions f_2 (U) and f_3 (U) defining the dependencies of the strength and cost of the cocktail on its composition U (the symbol Ω_1 designates the set {U} satisfying the constraints on strength and cost).

The symbols for optimization are introduced here for the first time; they will be used further widely (the general statement of the problem of optimization and effective methods to solve it are given in the 'Supplement').

In case the guests are bothered more with the strength

of the cocktail, the problem is as follows:

$$f_2(\mathbf{U}) \to \max_{\mathbf{U}\in\Omega_2}, \quad \Omega_2: \begin{cases} f_1(\mathbf{U}) \geqslant a_1, \\ f_3(\mathbf{U}) \leqslant a_3, \end{cases}$$

i.e. by selecting a certain composition U you should ($\to$) maximize (max) strength keeping its taste and cost within the stated limits (a_1 and a_3).

And finally, for the host hard pressed for the wherewithal, it is natural to state the problem as:

$$f_3(\mathbf{U}) \to \min_{\mathbf{U}\in\Omega_3}, \quad \Omega_3: \begin{cases} f_1(\mathbf{U}) \geqslant a_1, \\ f_2(\mathbf{U}) \leqslant a_2, \end{cases}$$

i.e. varying the composition U you should ($\to$) minimize (min) the cost of the cocktail meeting certain requirements Ω_3 to its taste and strength at the same time.

Then, in compliance with § 1.4 and the diagram of Fig. I.4.1, follows stage 2, isolation of the object of control from the medium. But the object, cocktail, has been predetermined, the guests being the sensors.

The next (third) stage is the structural synthesis of the model of the object. You are supposed to determine the nature of controlled and uncontrolled effects on the object and then the structure of the model, i.e. the type of dependencies of the properties of your cocktail on the selected alcoholic and other beverages and on their relative contents.

Therefore, in the process of structural synthesis of the model of the object, it is necessary to define the form of the functions f_j (U) ($j = 1, 2, 3$). First, you should determine the structure of control, i.e. the form of the vector $\mathbf{U} = (u_1, \ldots, u_i, \ldots, u_n)$. In other words, you should define every variable qualitatively, i.e. indicate what juice or wine 'hides' behind the index i in the variable u_i, and to determine the dimension n on the vector U of the composition. Then you should find the form of the dependence of the pertinent functions on the argument U.

The strength and cost of the cocktail are, as is known, strictly linear functions of the concentration of the ingredients:

$$f_2(\mathbf{U}) = \sum_{i=1}^{n} c_{i2}u_i, \quad f_3(\mathbf{U}) = \sum_{i=1}^{n} c_{i3}u_i,$$

where c_{i2} is the relative strength of the i-th ingredient, c_{i3} is the cost of a litre of the same. To define these parameters, suffice it to have a look at the bottles' labels and price tags. For instance, if the indicated strength of a dry wine is 10°, then the relative strength $c_{22} = 0.1$.

Now what is the function of the taste f_1 (U)? How much does the taste of the cocktail depend on the ratio of its ingredients? Evidently, it is going to be different for each of the guests. But since this is a common cocktail party, and the time is limited, i.e. there is no time to investigate into the taste of each guest, the function f_1 should be set plain enough. The simplest is the linear function

$$f_1(\mathrm{U}) = \sum_{i=1}^{n} c_{i1} u_i,$$

where c_{i1} are the 'taste' coefficients of the quality of the i-th ingredients. No doubt, the linear function reflects the real dependence of taste on the contents of the ingredients only approximately. However, a happy choice of the coefficients c_{i1} ($i = 1, \ldots, n$) might help approximate the function well. That being done, the structural synthesis of the model of the controlled object is completed.

It is necessary now to determine the numerical values of the 'taste' coefficients c_{i1} ($i = 1, \ldots, n$). First, there are no objective reasons for their determination, and there may be none, since it is well known: 'every man to his own taste'. In order to make evaluation of those coefficients, you come over to the stage of identification. You have to use the data obtained from observations over your guests in similar circumstances. In case there are no such data, you should get down to the stage of experiment design. There are many ways to carry it out. The easiest one is to ask your guests what they will prefer. The would-be consumer of the cocktail should be asked a number of questions so that the answers would give data suitable for the determination of the sought coefficients. This interrogation is an experiment with the object, the difficulty and complexity of which is defined by the number of questions and their difficulty. And if you have in your bar some juices and wines your guests are not familiar with, the experiment should include their tasting. Having

carried out proper experiments (how to do that, we are going to discuss in Chapter 3) with the consumers of the projected cocktail, you determine the 'taste' coefficients, and that will be the end of the object's model synthesis.

Now comes the stage of decision-making, or the synthesis of control (to be more precise), i.e. the determination of the optimal composition U^*. Let us hope you also elect to aim at gourmets, so the problem is (I.5.1), where the functions f_j ($j = 1, 2, 3$) are already known.

For instance, let us consider a case of three ingredients ($n = 3$): $i = 1$ is fruit juice, $i = 2$ is dry wine, and $i = 3$ is cognac. Let us suppose the data c_{ij} for them are:

	$i=1$ (juice)	$i=2$ (dry wine)	$i=3$ (cognac)	Source of information
$j=1$ (taste)	$c_{11}=1$	$c_{21}=0.75$	$c_{31}=0.5$	opinion
$j=2$ (relative strength)	$c_{12}=0$	$c_{22}=0.1$	$c_{32}=0.4$	label
$j=3$ (cost per litre in roubles)	$c_{13}=1$	$c_{23}=3$	$c_{33}=16$	price tag

To make further argumentation definite, let us define the objective. Let $a_2 = 0.1$ be the minimal tolerable strength of the cocktail, and $a_3 = 3.5$ be the maximal affordable cost per litre (these are the constants of the objective aimed at gourmets). Now the problem to synthesize the optimal composition of the cocktail may be put down as:

$$u_1 + 0.75u_2 + 0.5u_3 \to \max_{u_1, u_2, u_3 \in \Omega_1},$$

$$\Omega_1: \begin{cases} 0.1u_2 + 0.4u_3 \geqslant 0.1, \\ u_1 + 3u_2 + 16u_3 \leqslant 3.5, \\ u_1 + u_2 + u_3 = 1, \\ u_i \geqslant 0 \quad (i = 1, 2, 3), \end{cases}$$

where the constraints Ω_1 also include some obvious conditions for u_i:

$$\sum_{i=1}^{n} u_i = 1 \text{ and } u_i \geqslant 0.$$

The third condition of Ω_1 allows us to exclude $u_3 = 1 - u_1 - u_2$. Obvious transformations of the problem

lead to:

$$0.5u_1 + 0.25u_2 + 0.25 \to \max_{u_1,\, u_2 \in \Omega_1'},$$

$$\Omega_1': \begin{cases} 4u_1 + 3u_2 \leqslant 3, \\ 15u_1 + 13u_2 \geqslant 12.5, \\ u_1 + u_2 \leqslant 1, \\ u_1 \geqslant 0,\ u_2 \geqslant 0. \end{cases}$$

The area Ω_1' is shown in Fig. I.5.2 (it is within the cross-hatched lines). The area is the triangle ABC in the upper left corner. Now, we should find in the triangle a point maximizing the linear function

$$z_1 = 0.5\, u_1 + 0.25\, u_2 + 0.5$$

giving the taste properties of the cocktail. Evidently, the linear function is maximal somewhere on the boundary of the area Ω_1': more precisely, on one of the apexes of the triangle ABC. Let us calculate the values of z_1 in the apexes:

$$z_1\,(\mathbf{A}) = 0.75;$$
$$z_2\,(\mathbf{B}) = 0.5 \cdot 0.21 + 0.25 \cdot 0.75 + 0.5 = 0.78;$$
$$z_3\,(\mathbf{C}) = 0.25 \cdot 0.94 + 0.5 = 0.74.$$

Therefore, the point **B** is the solution of the problem. Consequently, the optimal composition of the cocktail in the first approximation is

$$u_1^* = 0.21; \quad u_2^* = 0.72; \quad u_3^* = 0.07,$$

where u_3^* is calculated from $u_3 = 1 - u_1 - u_2$. The taste of the cocktail is defined by the value $z_1^* = 0.78$. The recipe should be considered the first optimal solution. And this is the end of the control synthesis.

Let us skip the stage of the control realization. It is hardly troublesome. However, if you fail to find the required wines at the nearest shopping centre (the medium) that may lead to a reconsideration of the model's structure, i.e. it may make you pick some other ingredients for your cocktail. And that is going to be precisely the correction of the control due to a change in the medium

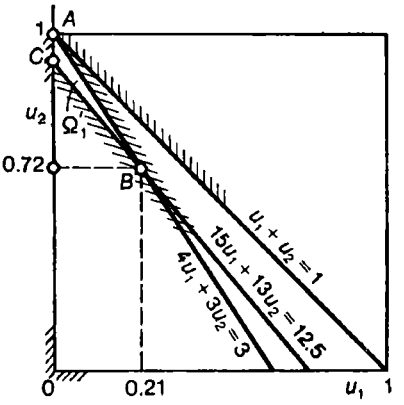

Fig. I.5.2. The plane of cocktail composition with $a_2 = 0.1$, $a_3 = 3.5$

by the moment of the control realization. This kind of correction is necessary for reaching the objective of control effectively.

Let us come over to the next stage, correction. The response Y of your guests to the realized control (the cocktail) ascertains the validity (adequacy) of the model. You can hardly expect unanimous approval of your recipe. There will be some disagreement in estimation. However, the average estimate, if the model is correct, should be the same as calculated, i.e. $z_1^* = 0.78$. If the average expert rating differs from that considerably (even in case the departure is to the better), the model requires correction, i.e. a change in the parameters c_{i1} ($i = 1, \ldots, n$). For instance, if an expert rating of the cocktail is $z_1 = 0.6$, the parameters of the model may be corrected as follows:

$$c'_{i1} = (0.6/0.78)\, c_{i1} = 0.77 c_{i1},\ i = 1, \ldots, n,$$

where c'_{i1} are new values of the parameters, yielding $z_1^* = 0.6$. Your next cocktail on the basis of the corrected model is going to taste better for the guests.

In case such correction does not improve the control, you should correct the structure of the model, e.g. assuming a non-linear function of the taste of the cocktail, or taking other ingredients, etc.

Somehow or other, the data obtained after the control has been realized are initial for the correction of the controlled object's model. And if the object, as in the cocktail example, is a complex system with the properties varying

in time due to a drift of its characteristics, correction of the model is unavoidable even in the hypothetical case of accurate synthesis of the model at some initial moment.

The correction may reach the object of control if you apply to producers with a request to make new kinds of juices and wines. And, after all, you may correct the objectives changing the parameters a_1, a_2, a_3 and switching over to another extremized criterion (yielding coffee or cucumber pickle!).

In conclusion, let us note that the cocktail example is supposed to be only a joking presentation of the stages of the complex object control. In case the reader is going to try and use the approach for making a cocktail in earnest, the author pleads not guilty for the consequences.

Comments on References to Introduction

The first books on cybernetics written by N. Wiener almost thirty years ago have not lost their grip so far. Being a classic, N. Wiener was able to put the fine and deep aspects of the new science into clear-cut words. Therefore it will be interesting for the reader to read the books [1, 2].

There are many accessible books on cybernetics. Among them are especially notable the one by W. Ashby [3], in which easy style is amazingly combined with in-depth analysis, and the book by St. Beer [4] who managed to formulate and solve without any mathematics a number of challenging problems of control for such complex objects as modern industrial plants.

Popular-science publications on cybernetics usually cover various aspects of the field [5-7]. The first book on cybernetics in the Russian language [5] should be noted.

The eight stages of the complex object control, described in § 1.4, are the concrete instructions to be applied if necessary. However, the control of a complex object may be realized only in case it appears possible to formalize both the object and the process of control. When such formal approach is too cumbersome or impossible, one should try the so-called systems approach (or analysis) which, as a matter of fact, is a kind of semi-formalized common sense. There are many books and manuals in the line. Let us note the basic ones.

The stages of system analysis (e.g. set forth by V. M. Glushkov [8]) are often the same as the stages of the complex object control mentioned in § 1.4. The highlights of the approach are very clearly presented in the book [9] which also compares its various modifications described in [10-13]. Appealing generalizations of the systems approach may be found in the books [14-16].

The systems approach in the complex object control is commonly tied up with the management of the objects, i.e. with informal

methods of control. That is well presented in [15]. And that is why the systems approach was not tackled with in this book which is engaged in the formal methods of control rather than management (see the 'Foreword').

From the viewpoint of the systems approach (or analysis), the eight-stage procedure of the complex object control put forward in this book may be considered as some sort of applied systems analysis, being a collection of formal methods (combined with using informal sources of information—experts) for constructive solution of the problem of the complex object control.

References

1. Wiener, N. *Cybernetics, or Control and Communication in the Animal and the Machine.* Technology Press, Boston, Mass., 1948.
2. Wiener, N. *The Human Use of Human Beings: Cybernetics and Society.* Eyre and Spottiswoode, London, 1954.
3. Ashby, W. Ross. *An Introduction to Cybernetics.* Chapman and Hall, London, 1956.
4. Beer, St. *Cybernetics and Management.* Wiley, Inc., New York, 1959.
5. Poletaev, I. A. *The Signal.* Sov. radio, Moscow, 1958 (in Russian).
6. Rastrigin, L. A. *This Chancy, Chancy, Chancy World.* Mir Publishers, Moscow, 1973.
7. Rastrigin, L. A., Grave, P. S. *Cybernetics as It Is.* Molodaya gvardia, Moscow, 1975 (in Russian).
8. Glushkov, V. M. *Introduction to ACS* (automated control systems). Tekhnika, Kiev, 1974 (in Russian).
9. Chernyak, Yu. I. *Systems Analysis in the Management of Economy.* Ekonomika, Moscow, 1975 (in Russian).
10. Optner, S. L. *Systems Analysis for Business and Industrial Problem Solving.* Prentice-Hall, Englewood Cliffs, N.J., 1970.
11. Fedorenko, N.P. On the methods of socio-economic prognostication. In: *Methodology of Prognostication of the Economic Development of the USSR.* Ekonomika, Moscow, 1971 (in Russian).
12. Young, S. *Management: a systems analysis.* Scott, Foresman and Co., Glenview, Ill., 1966.
13. Cleland, D. I., King, W. R. *Systems Analysis and Project Management.* McGraw-Hill Book Co., New York, 1968.
14. *Foundations of the systems approach and their application to the development of territorial automated systems of control.* Tomsk University Press, Tomsk, 1976 (in Russian).
15. Johnson, R., Kast, F., Rosenzweig, J. *The Theory and Management of Systems.* McGraw-Hill Book Co., New York, 1967.
16. Peregudov, F. I. (ed.). *Systems Design of the ACS for the Economy of a Region.* Statistika, Moscow, 1977 (in Russian).

Chapter 1 Determination of the Objectives of Control

Every kind of control is underlain by information about what the subject aims at, what he wants from the object, and what his objectives $\{Z^*\}$ are (strictly speaking, his set of objectives). Precisely this information defines both the object and its control system. The process of determination of the objectives comes before the whole procedure of control and therefore shapes it considerably.

1.1. The Paradox of the Objective

Generally, there is a bilateral link between the objective and the object in which the objective is realized: the objective bears upon delineation of the object, and the object in its turn determines the core of the objective. This mutual dependence of the objective and the object is due to that the subject cannot determine his objective without at least some preliminary (a priori) concept of the object in which the objective has to be realized. The audit of the fact whether the objective has been reached or not is carried out by way of analysis of the state Y of the object. Therefore, to check the realization of the objective, one is bound to know what to observe (to measure) in the object. This means that one should have already in the phase of determination of the objectives a fair enough concept of the object of the projected control.

Thus, there exists a peculiar paradox: the objective is determined by the object, and the object by the objective. The solution of the paradox is rather simple. The subject, while determining the objective, always has some concept of the object. The concept may be very approximate, but always reflecting some of its features, which are enough to determine the objectives of control. And that is why the determination of the objective with a certain degree of accuracy is almost always possible.

The object of control is completely determined by the objectives of control. However, one should bear in mind that far from every object satisfies the objectives of control and only occasionally are they fully satisfied, i.e. few of the formulated objectives can be attained given the resources allocated for control. As a matter of fact, it is always possible to formulate unreal objectives, such that cannot be realized with limited resources in any object. Usually, this compels to correct the objectives of control, which in its turn unavoidably changes the object of control, and so on. That is why the objective does not appear separately from the object. Apparently, the subject thinks over different versions of the objective and the object in parallel, they are mutually linked and lively interact. However, methodologically, it is more convenient to distinguish the pertinent processes, i.e. to consider the synthesis of the objective independently from the phase of separating the object from the medium, though in fact they are simultaneous.

Consequently, the above-mentioned paradox of the objective may be solved conveniently when the feed-back effect of the object on the objective is excluded. This effect may be taken into account at the next stage, when the object has been determined and its possibilities studied from the point of view of reaching the set objective. In case it turns out that there is no object satisfying the set objective well enough, it is necessary to correct the objective, since it cannot be reached then.

Now let us consider the process of determination of the objectives and study its specific features. It is quite in order to recall here that the objectives discussed are put forward by the subject. It is natural, therefore, to try and look into what the subject is whose objectives we suppose to realize by means of the synthesized control system. We need to consider the model of the subject for that.

1.2. The Model of the Subject

Let us note first of all that the subject is an active part and reflects during his interaction with the medium not only the state of the medium but also his own interests and needs due to his living activity. Let us note that

when a large and complex object is controlled, the subject may be not only a single person, but also a large group of people. Therefore, generally speaking, we shall mean by a subject an active system whose interests are to be reflected in the control system (for the sake of which it is to be produced).

Suppose the subject has k different needs, every one of which we shall characterize by number α_i, the degree of its manifestation, urgency, actuality. Evidently, the needs change in time and depending on the stage of the subject and its medium. Consequently, in each time moment the subject is characterized by a set (vector) of his needs:

$$A = (\alpha_1, \ldots, \alpha_i, \ldots, \alpha_k). \tag{1.2.1}$$

It is natural to assume that

$$\alpha_i \geqslant 0, \quad i = 1, \ldots, k, \tag{1.2.2}$$

i.e. an urgency of a need may not be negative, and when there is no i-th need, $\alpha_i = 0$. For instance, in the simplest case, we may consider only three primitive needs: thirst, hunger, and sex ($k = 3$). Obviously, such a subject is going to produce only 'animal' objectives that will define his behaviour.

Let us look forward and assume that the subject is capable of reaching the set objectives, i.e. he has a system to realize the objectives (that is precisely the control system). Then his behaviour will add up to the determination of the objective of the control and realization of the set objective, i.e. to carrying out the relationship

$$Z = Z^* \tag{1.2.3}$$

by means of the control U.

Let us recall that Z is the state of the object described in the language of the objectives $\{Z^*\}$. The 'translation' from one language into another is the function Ψ (see (I.1.2)): $Z = \Psi(Y)$, where $\Psi(\cdot)$ is a given function determined on the states Y of the object. We shall not consider how the subject manages to realize the objectives of control, i.e. how he produces U. Suffice it to know that somehow he can do that, i.e. he has an effective control system at his disposal, which 'follows' the objectives Z^* given to it.

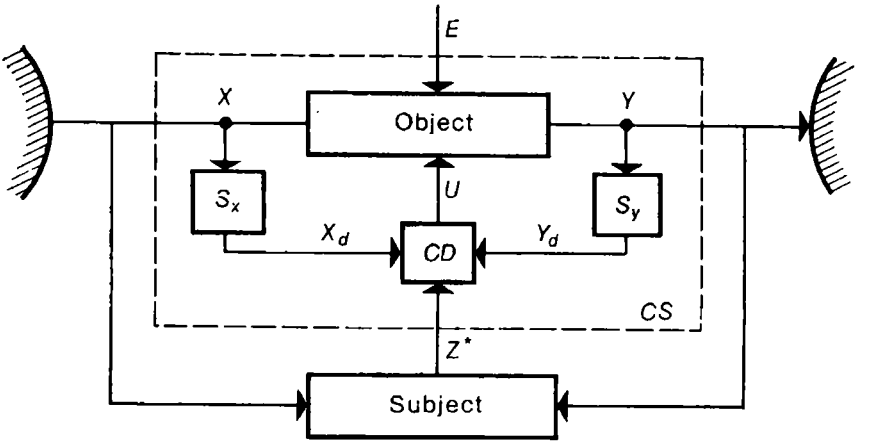

Fig. 1.2.1. Interaction between the subject and the control system

The relations between the subject and the control system (CS) are shown in Fig. 1.2.1, where the dashed line shows the CS for the realization of the objective Z^* given to it. Here the subject is the producer of the objectives to be followed by the CS. The needs of the subject are satisfied through the channel Y.

The problem of the subject, therefore, is reduced to the effective determination of the objectives of control. For that, he has, above all, to distinguish the 'good' objectives from 'bad' ones. Mathematically, the subject has to have a special function $\mu(Z^*)$ of the measure of quality of the objective, i.e. a special function of discomfort determined on the whole set $\{Z^*\}$ of all possible objectives. The function allows comparing two objectives Z_1^* and Z_2^* in the following way. If

$$\mu(Z_1^*) < \mu(Z_2^*) \tag{1.2.4}$$

then the objective Z_1^* is preferable to ('better' than) the objective Z_2^*, which is expressed as

$$Z_1^* \succ Z_2^*, \tag{1.2.5}$$

where $\succ$ is the symbol of preference (to be read 'preferable to'). When $\mu(Z_1^*) = \mu(Z_2^*)$, the objectives are equivalent, i.e. $Z_1^* \sim Z_2^*$. Let us note that the objectives may be equivalent for the subject, but by no means the same.

Consequently, the search for an optimal objective is

reduced to minimization of the function $\mu(Z^*)$, i.e. to the solution of the problem

$$\mu(Z^*) \to \min_{Z^* \in \{Z^*\}} \Rightarrow Z^{**}. \tag{1.2.6}$$

This expression means: one should ($\to$) minimize (min) the function $\mu(\cdot)$, by varying its argument Z^* within the limits of the given set of the objectives $\{Z^*\}$ which consists of the admissible objectives. The result that follows (this is denoted with the symbol $\Rightarrow$) is Z^{**}, the best objective in the circumstances.

Let us give some examples of producing the measure $\mu(\cdot)$. It may be defined as a weighted sum

$$\mu = \sum_{i=1}^{k} b_i \alpha_i, \tag{1.2.7}$$

where $b_i > 0$ $(i = 1, \ldots, k)$ are the 'weights' of the needs, characterizing the significance of the corresponding need to the activity of the subject. These values may be determined by means of expert evaluation, i.e. through a properly organized opinion poll of some experts knowing the needs of the subject well (see § 2.4).

One may assume another relationship

$$\mu = \max_{i=1, \ldots, k} b_i \alpha_i, \tag{1.2.8}$$

where μ is characterized by the most important need.

It is easy to note that the value of μ shows discomfort of the subject. It is natural for him to want to make it minimal by means of selecting an appropriate corresponding objectives Z^*, the realization of which would decrease that discomfort most.

The realization of the selected objective should change the needs of the subject towards their reduction (if the objective is 'good'):

$$A_1 \xrightarrow{Z^*} A_2, \tag{1.2.9}$$

where A_1 are the initial, and A_2 the resulting from the realization of the objective Z^* needs (1.2.1). It is natural to demand that A_2 minimize the function (1.2.7) or (1.2.8) at $b_i = \text{const}$. But it is necessary for that to establish the relationship (1.2.9) in the form of dependence

$$A_2 = \psi(A_1, Z^*, S), \qquad (1.2.10)$$

where S is the situation in the medium for the moment of selecting the objective Z^*; in the trivial case $S = \langle X, Y \rangle$, i.e. just the state of the medium around the subject.

If that dependence were known, then substituting the data into (1.2.7) or (1.2.8), we could obtain the sought function $\mu(Z^*)$, the minimization of which would define the optimal objective of the subject unambiguously. However, the relationship (1.2.10) may be realized so far by the musing of the following kind: 'My needs are A_1. In the situation S if I choose the objective Z^* and reach it, my needs will become A_2. If A_2 satisfy me, I choose that objective. If not, I try another objective', and so on. So far, we can only suppose the way the subject defines the state of his needs A_2 by the initial states of the situation S, his needs A_1, and the objective Z^*. This is when it helps to use the models of the medium and his own person, and to apply the experience of the use of such models.

Thus, the above mechanism for determination of the optimal objectives of control may underlie the construction of a formal 'producer of objectives'. For that we should have two functions: the function of discomfort $\mu(A)$ in the form (1.2.7) or (1.2.8) or any other, and the function of the change of needs $\psi(\cdot, \cdot, \cdot)$ (1.2.10), the latter making it possible to determine the needs resulting from the realization of the objective Z^*. So far only the human being can cope with the problem fully. But that does not mean that the process of the determination of the objective should be left by itself. There are concrete means to express the objectives, and if the objective is on the conscious level, it should be expressed with definite terms forming the language for the determination of the objectives Z^*. So let us consider that language, i.e. let us delineate the way to put the objectives Z^* formally.

1.3. The Spaces of Situations and the Objectives

The subject in the process of his communication with the surrounding medium focuses primarily on its parameters (properties) $s_1, \ldots, s_l$, that effect his needs, i.e. the values α_i $(i = 1, \ldots, k)$, defining the discomfort of the subject. By the values of these parameters the subject may evaluate the state of the medium. But being active, the subject naturally wants to change the medium so that to lessen his discomfort, i.e. to control it. This is what makes him look for such parameters in the medium which he is able to change so that to reduce his discomfort in the medium (let us remind that there is no object so far, it is to be isolated from the medium yet, from that part of it which the subject would like to change). It is natural to assume that the subject, while forming his objectives, takes into account only those parameters which can help improve his state in the medium. Thus, for instance, a driver, trying to increase the speed of the car, pays attention only to two factors of his medium: the position of the foot throttle and the readings of the speedometer. The brakes, clutch, and other factors do not bother him then.

Generally speaking, the parameters, which define the needs of the subject but cannot be changed by him, indirectly effect his behaviour in determining the objectives. Now here is when the mechanism of emotions 'plays', forming a mood (irritation, for instance), which cannot help effecting the process. Thus, the driver's unsatisfied desire to overtake other cars while the engine power is insufficient is bound to cause nervousness and effects, in the long run, the driver's behaviour. But the mechanisms of that kind are not going to be considered here.

Thus, the subject feels the surrounding medium as a certain situation that may be described as a finite or infinite set of parameters

$$S = (s_1, \ldots, s_l), \tag{1.3.1}$$

each of them 'worrying' the subject, since it effects his discomfort, and may be changed, i.e. as the situation

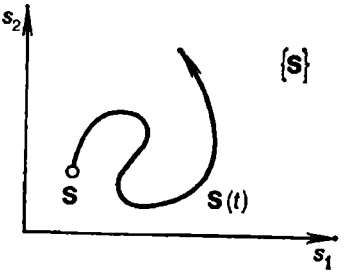

Fig. 1.3.1. Trajectory of the drift (evolution) of the situation S (t) in the space of situations

which is always controllable:

$$\mathbf{S}(\mathbf{U}) = [s_1(\mathbf{U}), \ldots, s_l(\mathbf{U})], \qquad (1.3.2)$$

where $\mathbf{U} = (u_1, \ldots, u_q)$ is the subject's control, i.e. $u_1, \ldots, u_q$ are the controllable parameters of the medium.

For instance, the situation parameters to be checked by a driver are: s_1, the speed of the car; s_2, the position (turn) of the wheels; etc., while the controllable parameters are: u_1, the position of the foot throttle; u_2, the position of the steering wheel; etc.

Now let us introduce the concept of the space of the situations {S}. It is formed by the above-mentioned parameters s_i ($i = 1, \ldots, l$) (1.3.1). Each point in that space defines a concrete situation around the subject (Fig. 1.3.1). The space {S} reflects both the medium surrounding the subject and various objects. The natural drift of the situation caused by the change (evolution) of the medium and possibly related to the actions of other subjects results in the displacement of the point S along some trajectory (see Fig. 1.3.1).

Therefore, $s_1, \ldots, s_l$ are the parameters of the medium percepted by the subject that effect him, and may be changed by means of certain actions U upon the medium.

But the subject formulates his objectives not in terms of the state of the medium S. It is more convenient for him to operate with other notions of his own that may be presented as variables z_i ($i = 1, \ldots, p$) (let us call them the objective variables). These variables may be described by a vector

$$\mathbf{Z} = (z_1, \ldots, z_p), \qquad (1.3.3)$$

where each objective variable z_i is unambiguously defined by the situation S, i.e.

$$z_i = \psi_i (S), \quad i = 1, \ldots, p, \tag{1.3.4}$$

and the functions $\psi_i (\cdot)$ determine the relationship between the state S of the medium and that of the objective variable z_i. At each concrete moment, z_i is a number characterizing the value of this objective parameter (e.g. temperature, length, cost, etc.), or shows the presence (1) or absence (0) of a certain property of the medium (e.g. some factors important for the subject). This relationship is expressed in the vector form as follows:

$$\mathbf{Z} = \Psi (S), \tag{1.3.5}$$

where

$$\Psi (\cdot) = (\psi_1 (\cdot), \ldots, \psi_p (\cdot)) \tag{1.3.6}$$

is a certain given vector-function linking the state of the medium and the objective variables.

The transformation Ψ of information into the form Z is also necessary because the subject usually defines his objectives in terms and notions associated with but not identical to the measured ones. In a particular case, it may be that $\mathbf{Z} = \mathbf{Y}$ $(p = l)$, but that is uncommon. For instance, when regulating the temperature of the object, suffice it to measure the temperature T, i.e. $z = y = T$. But to make the conditions comfortable, the operator has to measure not only the temperature, but also humidity, while the objective may be determined as a limitation on their certain combination.

Let us consider a p-dimensional space of the objectives $\{\mathbf{Z}\}$, which is formed by points (1.3.3). This space is convenient for the subject, since he may express a requirement (an objective) related to each coordinate, the fulfilment of which, in the opinion of the subject, would result in meeting one or several of his needs A.

The subject determines his objective as a conditional vector of the objective

$$\mathbf{Z}^* = (z_1^*, \ldots, z_p^*), \tag{1.3.7}$$

where z_i^* are not numbers, but requirements for the state S of the medium, the requirements being expressed by means of objective variables $z_i = \psi_i (S)$. These objectives-

requirements may be of various nature, but the form of their presentation should be unified. They should be reduced to one of the following forms:

(1) z_i^* means to equalize: $z_i = a_i$; this is a requirement that the i-th objective variable $z_i = \psi_i$ (S) should be made equal to a given value a_i;

(2) z_j^* means to constrain: $z_j \geqslant b_j$; the requirement for the j-th objective variable not to be less than a given threshold b_j;

(3) z_l^* means to minimize: $z_l \to \min$, i.e. the l-th objective variable should be minimized.

Consequently, to set the objectives, it is necessary first to define their structure, i.e. their belonging to one of those three forms, and then to set the values of a_i and b_j ($i = 1, \ldots, k_1$; $j = 1, \ldots, k_2$), where k_1 and k_2 are the numbers of the objectives of the first and second kind. In case some of the objectives of the subject cannot be reduced to one of those forms, there cannot be formal creation of a control system to reach the objectives. But practice shows that the requirements 'to equalize', 'to constrain', and 'to minimize' may be applied to almost any situation that challenges the subject and for the control of which a control system is created.

Let us note that the objectives of the subject are first of amorphous, uncertain nature. At best, one may describe them in the language of preferences, i.e. indicate the most preferable situations which the subject strives for. The analysis of these situations makes it possible to create the objective vector (1.3.7).

Hence, the process of determining the objectives of the subject is linked with (1) the determination of a vector-function (1.3.6), and (2) delineation of the requirements for each component of that vector.

Example. There are m points inside a furnace where temperatures $s_1, \ldots, s_m$ are measured. The problem is to control the temperature rating inside the furnace (e.g. by changing fuel feed with time), so that the temperature drop between any two neighbouring points would not exceed a given value Δ.

The m-dimensional space of situations in the problem is formed by the readings $s_1, \ldots, s_m$ of the sensors, and the objective space {Z} has the dimension $(m - 1)$ and is defined by the functions

$$z_l = \psi_l (S) = | s_{l+1} - s_l |, \quad l = 1, \ldots, m - 1.$$

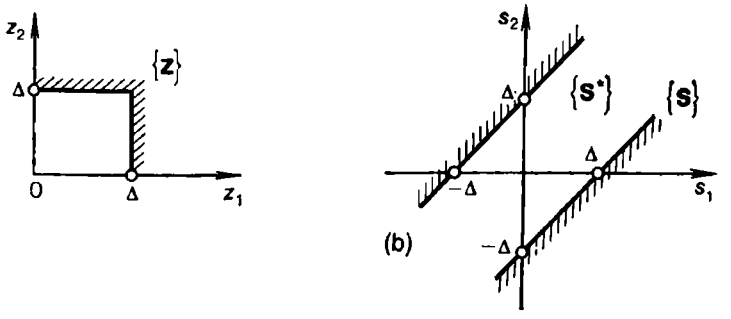

Fig. 1.3.2. Correspondence between the space of the objectives (*a*) and that of the situations (*b*) in furnace temperature control

The objective Z* (Fig. 1.3.2*a*) may be presented by the inequalities of the form

$$Z^*: z_i \leqslant \Delta, \quad i = 1, \ldots, m - 1 = k_3.$$

Let us consider how the objective Z* is reflected into the space of situations {S}. For that it suffices to consider the area S* defined by the system of all the objective requirements:

$$S^*: \begin{cases} \psi_i(S) = a_i, & i = 1, \ldots, k_1, \\ \psi_{k_1+j}(S) \geqslant b_j, & j = 1, \ldots, k_2, \\ \psi_{k_1+k_2+l}(S) \to \min, & l = 1, \ldots, k_3, \end{cases} \tag{1.3.8}$$

where k_3 is the number of the extremal objective requirements. The objective point or area S* which will satisfy all the requirements at the same time is the state of the medium the subject strives for. But whether or not he will manage that depends on his possibilities to influence the medium, i.e. on the type of dependence (1.3.2), and on the resources R at the subject's disposal. Constraints on the resources of control result in limitation of the control U for the subject

$$U \in \{U\}_R, \tag{1.3.9}$$

where $\{U\}_R$ is the set of control processes constrained by the energy, material, and other resources R of the subject.

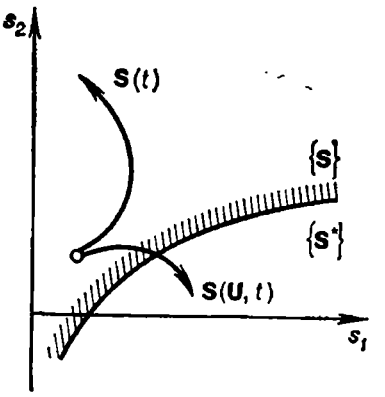

Fig. 1.3.3. Interaction of the trajectories of drift S (*t*) and control S (U, *t*) with the objective area {S*}

Now let us consider the objective area {S*} of the space of situations for the above example. It is determined by the inequalities

$$\{S^*\} : | s_{i+1} - s_i | \leqslant \Delta, \quad i = 1, \ldots, m-1.$$

The space of situations {S} including that objective area {S*} is presented in Fig. 1.3.2*b*. Evidently, the objective requirements Z* are more convenient to be shown in the objective space {Z}. Setting the objective area {S*} in {S} avoiding stage Z* is generally cumbersome.

Now let us consider the interaction of the objective area {S*} and the trajectory S (*t*) of the change in the medium under the impact of external factors (Fig. 1.3.3), i.e. the drift of the situation. In case the trajectory passes through the objective area, no control from the subject is needed. Suffice it to wait until the external influences cause S (*t*) $\in$ {S*}. However, it is unreasonable to count on such an unlikely though possible in principle coincidence. That is why the subject prefers to control the situation, i.e. he effects the medium purposefully and changes it:

$$\mathbf{S} = \mathbf{S}(\mathbf{U}, t), \tag{1.3.10}$$

to make

$$\mathbf{S}(\mathbf{U}, t) \subset \{\mathbf{S}^*\}, \tag{1.3.11}$$

i.e. to reach his objectives. Time *t* in these expressions denotes the effect of the drift on the properties of the medium.

Consequently, the subject needs the control U to reach his objective and compensate for the situation drift tending commonly to misfit the objective condition (1.3.11).

That is why every control should be considered from two points of view: both as a means to reach the sought-for objectives and as a means to compensate for unfavourable changes in the medium.

Let us note in conclusion that here and hereafter the parameters of a situation S in the medium are implicated to be measurable parameters of both the medium **X** proper and the object **Y**, the latter being changed by the subject by means of control **U**, i.e. $\mathbf{Y} = \mathbf{F}^{o}(\mathbf{X}, \mathbf{U})$. Therefore,

$$\mathbf{S} = \langle \mathbf{X}, \mathbf{F}^{o}(\mathbf{X}, \mathbf{U}) \rangle. \qquad (1.3.12)$$

However, S may thus be differentiated only after isolation of the object from the medium which is discussed in the following chapter.

Comments on References to Chapter 1

The primacy of the objective is often argued against since there is no objective in the amazingly complex, purposeful, and reliable biological systems of control. This paradox may be easily solved recalling that the objective in biological control systems is the state of another system higher up along the biological hierarchy. A similar principle is commonly applied in the 'man-made' systems of hierarchical control where the objective of each level is 'handed down' from the upper one. But that does not answer the problem to determine the objective by the upper level. The unsolved problem of the objective determination is redirected in the hierarchical set-up to the management. The management also has to cope with the specific task to coordinate the objectives put down to the subordinates. These problems are thoroughly dealt with in the book [1] (it seems to be the first detailed study of the hierarchical control systems), but the determination of the objective has not been considered there yet.

The model of the subject as described in § 1.2 is very superficial and cannot help formalize the process of the determination of the 'good' objectives. It is given here only to produce the impression of the existence of the complex mechanisms for the determination of the objectives which are employed by the subject and of which we do not know much yet. The model has been first described in [2].

In more detail the specificity of the subject as a purposeful system has been discussed in the fine-written book [3] where human behaviour has been tackled from the point of view of control as a system for reaching objectives, presenting an easily understandable in-depth analysis without excess formality. However, the mechanisms for the objective determination were not discussed there either.

Recently, very much attention to these mechanisms has been paid by psychologists who termed correctly the goal as the prime

basis of behaviour. And behaviour is traditionally a psychological domain. Hence in [4] the mechanisms of the goal formation are discussed from that viewpoint, but they are not coupled with the processes of control considered in the present book. The latter deals with formalization of the objectives, i.e. their numerical presentation.

References

1. Mesarovic, M.D., Macko, D., Takahara, Y. *Theory of Hierarchical Multilevel Systems.* Academic press, New York, London, 1970.
2. Grave, P.S., Rastrigin, L.A. On a mathematical model for the synthesis of behaviour (the subconscious level). In: *Adaptive systems.* Zinatne, Riga, 1972, 2, p. 5-15 (in Russian).
3. Ackoff, R.L., Emery, F.E. *On Purposeful Systems.* Aldine & Atherton, Chicago, New York, 1972.
4. *Psychological Mechanisms for Goal Formation.* Nauka, Moscow, 1977 (in Russian).

Chapter 2 Determination of the Object of Control

The problem of isolating the object from the medium arises after the determination and formal description of the set $\{Z^*\}$ of the objectives of control. The process of the isolation of the object should end in the determination of the boundaries of the object F^o (let us designate the object by its operator F^o) separating it from the medium, i.e., in such a description of the object which will define its boundaries unambiguously. There is no problem in a number of cases when the boundaries of the object are obvious. That occurs commonly when the object is independent enough (an aircraft, a ship, any device, a car, etc.). But there are other cases when the ties of the object with the medium are so strong and diverse that it is sometimes hard to see the boundary. That is precisely what is dealt with in this chapter.

2.1. Analysis of the Problem

It is natural to put a question whether it is important at all to mark the boundary dividing the object and the medium if it is nowhere evident. It turns out to be important, and very much so. Let us demonstrate this with an easy example.

Suppose we are going to create a system to control a technological process. Is the object delineated by the process alone? There is no answer to this question since the boundaries of the object of control depend on the objectives to be satisfied by the given process and its production. In case the objectives are clear enough (e.g. to reach a certain quality and smooth functioning), one cannot avoid managing not only the process itself, but its input as well, i.e. the raw material processed. That means that the object of control should include the sup-

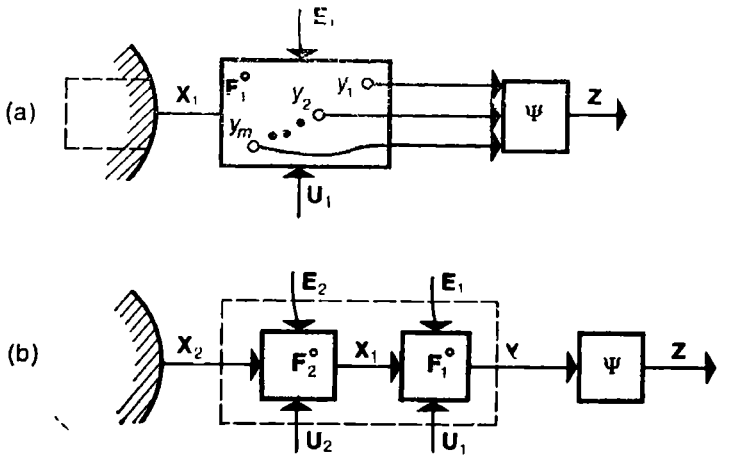

Fig. 2.1.1. To the comments on determination of the boundaries of the object

plier of the raw material as well. And sometimes there is a need to control the raw material delivered to the supplier. (Let us note that the need in such a control is one of the reasons for setting up industrial complexes, associations, amalgamations, etc., which are able to control production at every stage.)

But you can't go to the other extreme, i.e. try and control every single thing, though some formal grounds for that may be found always since the ties of various processes and phenomena may be followed far enough. However, there is a certain optimum of the 'size' of the object. Let us try and show it.

Two versions of the object of control are presented in Fig. 2.1.1. Here $y_1, \ldots, y_m$ are the points to obtain the information needed to evaluate the realization of the set of the objectives $\{\mathbf{Z}^*\}$ in the object; $\mathbf{Y} = (y_1, \ldots, y_m)$. The points should exist in any version of the object. The first version (Fig. 2.1.1*a*) presents the relation between the inputs and outputs as

$$\mathbf{Y} = \mathbf{F}_1^{\circ}(\mathbf{X}_1, \mathbf{U}_1, \mathbf{E}_1), \qquad (2.1.1)$$

where $\mathbf{X}_1$, $\mathbf{U}_1$, and $\mathbf{E}_1$ are the observed, controlled, and unobserved inputs respectively.

This version of the object may be expanded adjoining to it the part of the medium shown in Fig. 2.1.1*a* by the

dashed line. Trying to control this part of the medium, i.e. to consider it to be a part of the object, we can manage to effect the earlier uncontrolled input $\mathbf{X}_1$, which clearly extends the possibilities of control.

The dashed line in Fig. 2.1.1*b* shows thus expanded object for which

$$\mathbf{Y} = \mathbf{F}_1^o(\mathbf{F}_2^o(\mathbf{X}_2, \mathbf{U}_2, \mathbf{E}_2), \mathbf{U}_1, \mathbf{E}_1). \qquad (2.1.2)$$

The control $\mathbf{U}_1$ here is not excluded, it is added up with $\mathbf{U}_2$ which allows controlling the additional part of the object.

Clearly, the state **Y** in both versions is the same, but the media are different, and so are the effects of the media on that state. Which of the versions is better? To answer the question, one needs to weigh all the pros and cons of the two versions. Let us try it.

The first version (Fig. 2.1.1*a*)

Advantages:

—only one factor of uncertainty (unobserved input) $\mathbf{E}_1$,

—easier to find a mathematical description, i.e. the model of the object $\mathbf{F}_1^o$,

—lower number of the points of control $\mathbf{U}_1$ (less resources R are necessary for control).

Disadvantages:

—lower controllability, since the input $\mathbf{X}_1$ is uncontrollable (this factor has been decisive for the introduction of the competing second version of the object).

The second version (Fig. 2.1.1*b*)

Advantages:

—a possibility, if only partial, to control the input $\mathbf{X}_1$ which effects the state **Y** directly (the control is carried out through the additional channel $\mathbf{U}_2$).

Disadvantages:

—extended factor of uncertainty due to the additional input $\mathbf{E}_2$ of the new part,

—complication added to the mathematical description, since it is now necessary to describe the added part $\mathbf{F}_2^o$,

—increased number of the points of control; it expands the possibilities of control, but requires more resources R of control, and if they are unchanged, it results in lesser range of control $\mathbf{U} = \langle \mathbf{U}_1, \mathbf{U}_2 \rangle$, i.e. in lower possibilities of control.

It may seem at the first glance that there is too much disadvantage in the second version and it needs to give preference to the first. However, the only advantage of the second version may turn to be crucial. Our ultimate goal is to reach the established set of the objectives $\{Z^*\}$ best. Considering the attainability of the objectives, the second version may prove to be better than the first.

Consequently, the problem is to be able to distinguish well enough between 'good' and 'bad' versions of isolating objects from their medium. How shall we reach that? If we had a formal (mathematical) description of the medium, the process of isolation of the object from it would not present any difficulties in principle. Indeed, 'striking out' various 'pieces' from the medium and calling them various versions of the object, we might always check formally whether or not the established objectives of control were to be reached in the version. Repeating this procedure for various versions, we could find the version providing maximal (best of all a hundred per cent) controllability. That would be all!

But there is no formal description of the medium, and this approach is unacceptable. Therefore, we shall look for another approach. The direction of the search is rather obvious. Always, when there is no (at least for the time being) formal apparatus to solve a problem, the solution (though often a very approximate one) is sought from experts. This is the only possible way at the stage of the object determination. So let us discuss the process of expert participation in the solution of the given problem.

2.2. The Method of Expert Evaluation

We shall call 'an expert' a person whose opinion concerning our problem may be more competent than ours. Otherwise, there is no reason to apply to the person as an expert. Obviously, an expert in the problem of the object determination should:

—know well the medium from which the object is to be isolated;

—know well the objectives of control;

—have a good conception of possibilities and means of control;
—have a good conception of means for collection and processing the data on the medium and the object.

They do not ask an expert to determine the object. The expert is used as a source of information needed to make a decision on what should be considered the object of control. The decision is made by the designer of the control system. The expert is only presented with questions the answers to which contain some needed data. Far from all questions are advisable to ask even an expert (you may ask, but great care should be taken in considering the answers).

Let us start with the selection of the experts and their qualification. Let us estimate the qualification of each i-th expert, i.e. his competence in a j-th problem under discussion, by a so-called quotient of competence

$$0 \leqslant k_i^j \leqslant 1, \; i = 1, \ldots, N_j, \qquad (2.2.1)$$

where N_j is the number of experts invited to solve the j-th problem. When $k_i^j = 0$, the i-th expert is admitted to be incompetent in the j-th problem, while in case $k_i^j = 1$, the expert is recognized to be fully competent. If the quotients of competence are known, the number of experts may be clearly defined. In fact, the expert whose competence is below a certain threshold, i.e.

$$k_i^j < \delta^j, \qquad (2.2.2)$$

where δ^j is an established threshold of competence in the solution of j-th problem, may not be troubled with it. This is the way to define the number N_j of the experts to be invited to solve a j-th problem (the threshold δ^j is established by the designer).

Let us consider a method to define the competence quotients (2.2.1). The quotients are defined by the expert method as well, i.e. by means of mutual evaluation of competence of the experts themselves. This proceeds as follows. Every i-th expert, while solving a j-th problem, evaluates competence of each expert (and his own as well) by means of coefficient

$$0 \leqslant k_{il}^j \leqslant 1, \quad l = 1, \ldots, N_j. \qquad (2.2.3)$$

Here k^j_{il} is the opinion of the i-th expert on the competence of the l-th expert in the solution of the j-th problem (k^j_{ii} is the self-evaluation of competence of the i-th expert in the matter). Now, there is no difficulty in estimating the competence of the l-th expert as a mean value of all the evaluations including his own:

$$k^j_l = \frac{1}{N_j} \sum_{i=1}^{N_j} k^j_{il}, \quad l = 1, \ldots, N_j. \qquad (2.2.4)$$

Let us note that the matrix of mutual evaluations (2.2.3) may include some 'absent' dashes meaning that there is no evaluation (an expert does not know another and so cannot make proper evaluation). When this is the case, the mean value according to formula (2.2.4) is calculated by the existing evaluations, i.e. taking account of the dashes.

Here is an example of such a matrix for the competence evaluations:

i	l			
	1	2	3	4
1	1	—	0.3	0.5
2	0.7	0.8	0.4	—
3	—	0.8	0.7	0.6
4	0.7	0.7	—	0.9
k^j_l	0.80	0.77	0.47	0.67

In each of the cases the mean values k^j_l are estimated by the three existing evaluations.

Evidently, this method for experts' mutual evaluation does not take into account any possibly involved relationships between the experts as human beings. But that should be taken due regard of in reviewing such matrices of competence evaluations. Thus, if one of the experts gets very high estimates from some and very low esti-

mates from others, the chances are that there is either some peculiarity of his character or originality of his thought which is accepted by some, but rejected by the others. Recognition of the 'human' anomalies of this kind in a matrix of competence evaluation would help avoid the mistakes that occur when the opinion of seven companiable 'fools' becomes more important than the opinion of one unsociable 'sage'.

Having the values of competence quotients (2.2.4) ready, we may start solving the j-th problem. Every problem may be ultimately reduced to determination of either a figure, or a set of q figures which are called the solution to this problem. Let us consider the first case, since the second one usually boils down to a q-fold solution of the first one.

Suppose the problem is presented in the following manner: 'Proceeding from the conditions (the conditions are listed, an unknown factor W^j solving the problem among them) give your opinion on the value of this factor as a single number ($W_i^j = \ldots$). Thank you for your opinion!'

Having received the values W_i^j ($i = 1, \ldots, N_j$) from all the experts (some of them may refuse to give the opinion admitting their incompetence in the problem, and therefore they are to be excluded from N_j), one should make a decision on the value of the factor W^j. It is quite natural to determine it with due regard of the experts' competence as a weighted mean with the weights equal to the competence quotients

$$W^j = \left(\sum_{i=1}^{N_j} W_i^j k_i^j \right) \Big/ \sum_{i=1}^{N_j} k_i^j. \qquad (2.2.5)$$

It is clear that every expert's opinion is weighted with the value of his competence quotient. Therefore, the opinion of low-competent experts is automatically taken less account of. Consequently, the experts may not be chosen according to criterion (2.2.2). This criterion helps only not to bother the experts whose opinion is inessential for the solution of the j-th problem.

Formula (2.2.5) solves the problem. But one more important thing should be properly dealt with here. The

opinions of the experts, i.e. the values W_i^j may differ greatly from each other due to two reasons. On the one hand, the discrepancy may be due to that the question is hard to be answered. To make the task easier for the experts, the general set-up of the problem should be changed. On the other hand, it may be due to the existence, with regard to the factor W^j, of two contradicting viewpoints which should be taken into account in the solution of the problem. If this is the case, the designer should identify the viewpoints and adhere to one of them.

The diversity of opinions is easiest to make out by estimating the dispersion of the obtained experts' evaluations

$$\hat{D}(W^j) = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} \left(W_i^j - \frac{1}{N_j} \sum_{i=1}^{N_j} W_i^j \right)^2 . \tag{2.2.6}$$

Low value of this dispersion shows acceptable unanimity of the experts.

Clearly, the method of expert evaluation permits the solution of formally unsolvable problems. True, the solution thus obtained is always approximate, but it can be made more accurate by engaging more experts and considering their competence in the problem.

Now let us apply this method to the solution of the problem to isolate the object from the medium.

2.3. Controllability of the Object

It is natural that the procedure of isolation of the object from the medium should be constructed in such a way that the produced object would realize the given objectives $\{\mathbf{Z}^*\}$ of control, i.e. so that it could be effectively controlled in situations occurring in the process of control. Therefore let us introduce the concept of *controllability* of the object implicating the probability to reach the given objectives in various situations.

First, let us discuss the concept of the situation. In the most general case, the situation occurring in the

process of control should be understood as a set of three

$$\mathbf{S} = \langle \mathbf{X},\ \mathbf{E},\ \mathbf{Z}^* \rangle \qquad (2.3.1)$$

including the state of the uncontrollable inputs of the object and the objective to be reached (it is easy to see that this definition generalizes the one proposed above (1.3.12), where the control U depends in the long run only on X and Z*). All the situations to be met with in the process of control of an object may be subdivided into two subsets of situations: the controllable ones, when the objective Z* can always be reached, and the uncontrollable, when the objective Z* cannot be reached.

Let us denote by {S} the set of all possible situations S occurring in the process of control. Let $\{\mathbf{J}_{\bar{\mathbf{S}}}\}$ be the subset of situations {S} in which the object is uncontrollable, i.e., not all of the objectives {Z*} can be reached, and let $\{\mathbf{J}_{\mathbf{S}}\}$ be the subset of situations in which the object is controllable, i.e. all of the objectives {Z*} can be reached. Obviously

$$\{\mathbf{J}_{\mathbf{S}}\} \cup \{\mathbf{J}_{\bar{\mathbf{S}}}\} = \{\mathbf{S}\}. \qquad (2.3.2)$$

The expression implicates an evident thing: every situation S may be either controllable or uncontrollable.

Let us correlate every element of the set {S}, i.e. every situation S, with a value $\rho = \rho(\mathbf{S})$, the probability of emergence of this situation S. In case the number of elements of the set S is infinite, one should understand, as usual, by $\rho(\mathbf{S})$ the density of probability. In the first (discrete) case

$$\sum_{i=1}^{N} \rho(\mathbf{S}_i) = 1,$$

where N is the general number of the occurring situations. In the second (continuous) case

$$\int_{\{\mathbf{S}\}} \rho(\mathbf{S})\, d\mathbf{S} = 1, \qquad (2.3.3)$$

where the integral is taken over the entire set {S}. Relationship (2.3.3) presents the fact that no real situations S may occur beyond the area {S}.

The continuous case is more convenient and it generalizes the discrete case therefore let us use only this one, though for practical reasons we shall have to come back to the discrete case.

Now let us put more accurately the concept of controllability. Let us term as the *controllability of the object* the probability that a randomly chosen situation S is controllable. This probability

$$P = \int\limits_{\{J_S\}} \rho(S)\, dS, \tag{2.3.4}$$

where the integral is taken over the subset of the controllable situations $\{J_S\}$.

The uncontrollability may be defined similarly as the integral over the subset $\{J_{\bar{S}}\}$ of the uncontrollable situations

$$\overline{P} = \int\limits_{\{J_{\bar{S}}\}} \rho(S)\, dS. \tag{2.3.5}$$

It follows from (2.3.2) and (2.3.3) that

$$P + \overline{P} = 1. \tag{2.3.6}$$

If one of the considered versions of the object is more controllable, other conditions being equal, it should be preferred. Consequently, the problem to isolate the object is reduced to the problem to determine its controllability P.

However, any strict definition of the sets $\{J_S\}$ and $\{J_{\bar{S}}\}$, as well as of the function $\rho(S)$ is impossible, because there is no formal description of the medium. This is why experts are invited to solve this problem.

In order to determine (more accurately, to estimate) the controllability P of the object, it is necessary to decompose the problem, i.e. to present it as a number of simpler problems which may be solved by experts.

For that the experts should have the following initial data:

(1) the objectives of control, i.e. full enough description of the set of objectives $\{Z^*\}$ of control to be realized in the object;

(2) the data on the available resources R allocated to produce the control system and maintain it.

Let us designate these initial data by a pair

$$A = \langle \{\mathbf{Z}^*\},\ R \rangle. \tag{2.3.7}$$

The experts should be offered full information on the supposed object and its connections with the medium, i.e.:

(3) a description of the object $\mathbf{F}^o$,

(4) a description of the set $\{\mathbf{X}\}$ of the observed states of the medium which may be taken into account in the process of control synthesis (the 'volume' of this set apparently depends on the resources R allocated to set-up a system for measuring and monitoring $\mathbf{X}$),

(5) a description $\{\mathbf{U}\}_R$ of the set of controls allowed by the resources R (the available energy, time, memory store of the control device, the speed and features of data processing, etc.),

(6) a description $\{\mathbf{E}\}$ of the set of unobserved factors of the medium and the object.

Let us designate these data on the object by a set of four

$$B = \langle \mathbf{F}^o,\ \{\mathbf{X}\},\ \{\mathbf{U}\}_R,\ \{\mathbf{E}\} \rangle. \tag{2.3.8}$$

The initial data A and B being available, the experts should give estimates for the controllability P of the object. Let us denote by φ the algorithm according to which the experts operate, so that

$$P = \varphi(A, B). \tag{2.3.9}$$

However, it is easy to imagine that the job presented to the experts is hard. In fact, though the data A and B do contain the necessary information on the controllability P of the object, the experts cannot estimate the controllability P, i.e. realize the algorithm φ. The problem is too involved for them. Therefore we shall decompose the problem once again. Let us introduce for that various forms of the controllability and uncontrollability of the object.

We shall term the object *absolutely controllable* ($P = 1$), if every situation out of $\{\mathbf{S}\}$ is controllable, i.e. the objective is always achievable. This implies that for any observed state of the medium $\mathbf{X} \in \{\mathbf{X}\}$, for any unobserved

input $\mathbf{E} \in \{\mathbf{E}\}$, and for any objective $\mathbf{Z}^* \in \{\mathbf{Z}^*\}$ there always exists such a control $\mathbf{U}^* \in \{\mathbf{U}\}_R$ that will take the object into the required state, i.e.

$$\mathbf{Z} = \mathbf{Z}(\mathbf{X}, \mathbf{U}^*, \mathbf{E}) = \Psi(\mathbf{F}^o(\mathbf{X}, \mathbf{U}^*, \mathbf{E})) = \mathbf{Z}^*, \tag{2.3.10}$$

where $\Psi(\cdot)$ is the transformation of the space of states $\{\mathbf{Y}\}$ of the object into the space of objectives $\{\mathbf{Z}\}$ (see (1.3.6)); $\mathbf{Z}$ is the state of the object in the space of objectives; $\mathbf{Z}(\cdot, \cdot, \cdot)$ is the function showing the relationship of $\mathbf{Z}$ with $\mathbf{X}$, $\mathbf{U}$ and $\mathbf{E}$.

It is convenient for economic denotation of this definition to use the so-called quantifiers of existence ($\exists$) and generality ($\forall$). Then

$$\forall \mathbf{X} \in \{\mathbf{X}\}, \quad \forall \mathbf{E} \in \{\mathbf{E}\}, \quad \forall \mathbf{Z}^* \in \{\mathbf{Z}^*\},$$
$$\exists \mathbf{U}^* \in \{\mathbf{U}\}_R: \ \mathbf{Z}(\mathbf{X}, \mathbf{U}^*, \mathbf{E}) = \mathbf{Z}^*. \tag{2.3.11}$$

At first sight this is a complicated expression, but it can be read very easily: for any $\mathbf{X}$ belonging to the set $\{\mathbf{X}\}$ (which is denoted as $\forall \mathbf{X} \in \{\mathbf{X}\}$), and for any $\mathbf{E}$ belonging to the set $\{\mathbf{E}\}$ (i.e. $\forall \mathbf{E} \in \{\mathbf{E}\}$), and for any $\mathbf{Z}^*$ belonging to $\{\mathbf{Z}^*\}$ (i.e. $\forall \mathbf{Z}^* \in \{\mathbf{Z}^*\}$), there exists such $\mathbf{U}^*$ belonging to $\{\mathbf{U}\}_R$ (i.e. $\exists \mathbf{U}^* \in \{\mathbf{U}\}_R$) that $\mathbf{Z} = \mathbf{Z}^*$, i.e. the objective is achievable.

The absolute controllability ($P = 1$) is uncommon in the practice of complex object control. In fact, it requires either to considerably diminish the set of objectives or to have a constant medium ($\mathbf{X} = \text{const}$) or to allocate much greater resources R for control. As a rule, absolutely controllable are either simple objects (e.g. systems of automatic process control), or very important ones (e.g. life-support systems for operators in extremal conditions). The overwhelming majority of control systems for complex objects cannot be placed in this category.

Therefore, a concept of *partial* or *relative controllability* should be introduced. We shall look into the reasons of uncontrollability. Let us introduce various types of uncontrollability for the purpose. The X-uncontrollability is connected with the specificity of the set $\{\mathbf{X}\}$ where

$\mathbf{X}$ is found such that for any $\mathbf{E} \in \{\mathbf{E}\}$ there is no such $\mathbf{U} \in \{\mathbf{U}\}_R$ for which $\mathbf{Z} = \mathbf{Z}^*$, i.e. any objective is unachievable. Let us define the $\mathbf{X}$-uncontrollability formally

$$\mathbf{J}_{\overline{\mathbf{X}}}: (\exists \mathbf{X} \in \{\mathbf{X}\}: \forall \mathbf{E} \in \{\mathbf{E}\}, \forall \mathbf{Z}^* \in \{\mathbf{Z}^*\},$$

$$\overline{\exists} \mathbf{U} \in \{\mathbf{U}\}_R: \mathbf{Z}(\mathbf{X}, \mathbf{U}, \mathbf{E}) = \mathbf{Z}^*). \qquad (2.3.12)$$

Here $\mathbf{J}_{\overline{\mathbf{X}}}$ is the $\mathbf{X}$-uncontrollability and $\overline{\exists}$ is the nonexistence quantifier meaning 'there is no' ('cannot be found'). Expression (2.3.12) is to be read as follows: the $\mathbf{X}$-uncontrollability is defined as the existence of such a tolerable $\mathbf{X}$ that for any possible $\mathbf{E}$ and $\mathbf{Z}^*$ such that there is no possible $\mathbf{U}$ at which the objective $\mathbf{Z}^*$ is achievable.

Example. Let $\mathbf{X}^*$ be the energy required for normal operation of the object. Then when $\mathbf{X} < \mathbf{X}^*$, i.e. when the energy supply is insufficient, the object will not function and will not achieve the required objectives $\mathbf{Z}^*$ in spite of all the efforts $\mathbf{U}$ taken within the limits of the allocated resources R.

The $\mathbf{E}$-uncontrollability may be introduced in a similar manner as the existence of such unobserved states of the medium which do not allow with any possible $\mathbf{X}$ reaching any admissible objective:

$$\mathbf{J}_{\overline{\mathbf{E}}}: (\exists \mathbf{E} \in \{\mathbf{E}\}: \forall \mathbf{X} \in \{\mathbf{X}\},$$

$$\forall \mathbf{Z}^* \in \{\mathbf{Z}^*\}, \overline{\exists} \mathbf{U} \in \{\mathbf{U}\}_R: \mathbf{Z}(\mathbf{X}, \mathbf{U}, \mathbf{E}) = \mathbf{Z}^*). \qquad (2.3.13)$$

Example. Let $\mathbf{E}$ be the depreciation of the equipment showing in a decreased accuracy of meters. Obviously, there is a certain level of depreciation starting from which the drop in accuracy of the meters is going to be so large that it will not allow reaching with their aid the required objectives, e.g. turn out high-grade products even when special measures $\mathbf{U}$ are taken within the limits of the resources R.

The $\mathbf{Z}^*$-uncontrollability is characterized by such admissible objectives which are not always achievable given other possible conditions:

$$(\mathbf{J}_{\overline{\mathbf{Z}^*}}: (\exists \mathbf{Z}^* \in \{\mathbf{Z}^*\}: \forall \mathbf{X} \in \{\mathbf{X}\},$$

$$\forall \mathbf{E} \in \{\mathbf{E}\}, \overline{\exists} \mathbf{U} \in \{\mathbf{U}\}_R: \mathbf{Z}(\mathbf{X}, \mathbf{U}, \mathbf{E}) = \mathbf{Z}^*). \qquad (2.3.14)$$

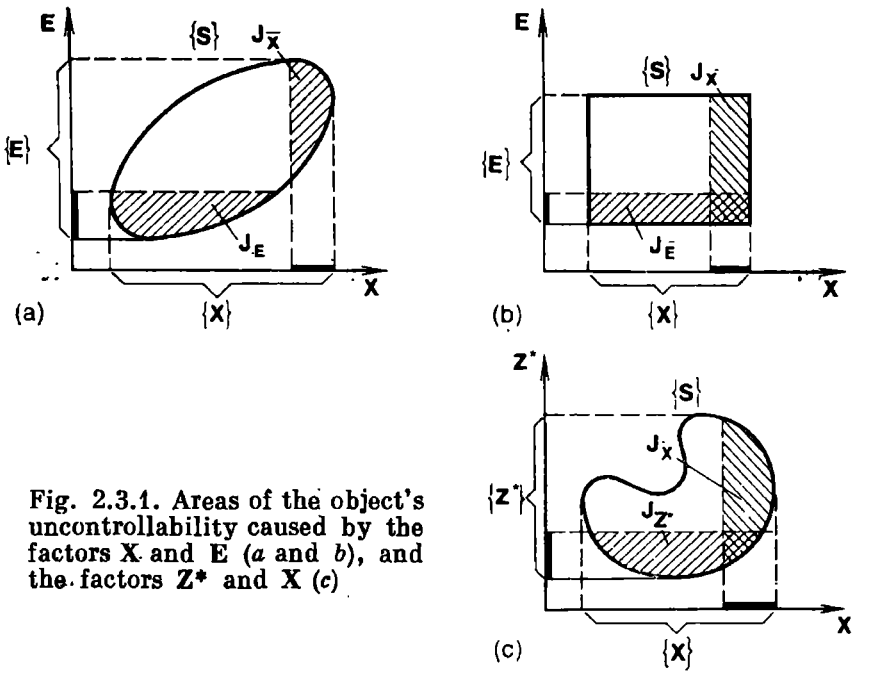

Fig. 2.3.1. Areas of the object's uncontrollability caused by the factors X and E (*a* and *b*), and the factors Z* and X (*c*)

This means that there exist some admissible objectives which are achievable not at all states of the medium. For instance, an objective to overcome the planned production target cannot be realized when there may be irregular supply (X) or an accidental failure (E) though both allow fulfilment of the plan.

Fig. 2.3.1 represents several cases of uncontrollability in the space of situations $\{S\} = \{X, E, Z^*\}$. The bold-type segments show the zones of uncontrollability that 'engender' areas of uncontrollabilities $J_{\bar{X}}$, $J_{\bar{E}}$ and $J_{\bar{Z}^*}$. Evidently, such zones may intersect (see Figs. 2.3.1*b*, *c*). Let us keep our mind on it, we are going to develop the idea further.

We shall introduce now the concepts of overlapping uncontrollabilities, i.e. those depending on two or three of the factors X, E, Z* at a time and bearing on the success of the control.

The XE-uncontrollability may be defined as a simul-

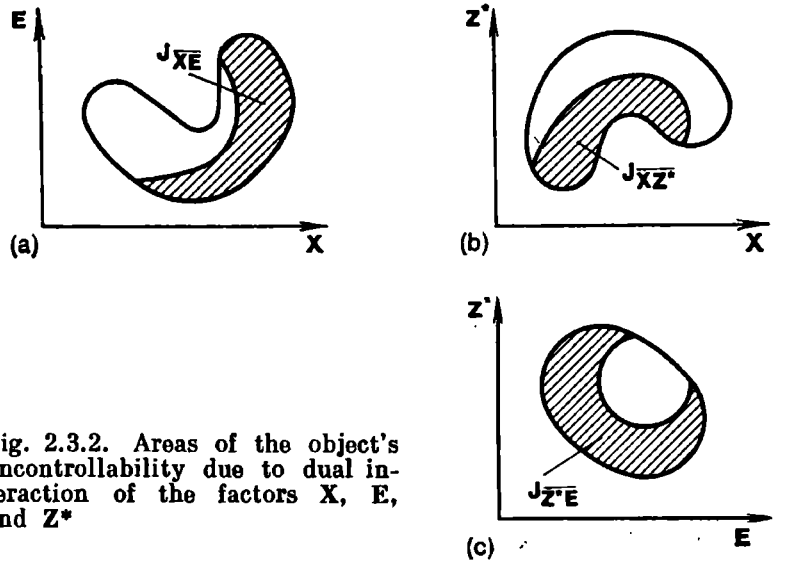

Fig. 2.3.2. Areas of the object's uncontrollability due to dual interaction of the factors X, E, and Z*

taneous existence of such X and E at which not every objective is achievable:

$$J_{\overline{XE}}: \ (\exists (X, E): \forall Z^*, \ \bar{\exists} U \in \{U\}_R : Z(X, U, E) = Z^*). \tag{2.3.15}$$

For instance, a simultaneous drop in the quality of raw material (X) together with depreciation (E) of the equipment may often jeopardize the fulfilment of the plan though each of the factors may be successfully dealt with by the appropriate control U.

Here and hereafter it is implied for shorter notation that $X \in \{X\}$, $E \in \{E\}$, $Z^* \in \{Z^*\}$, $U \in \{U\}_R$.

The XZ*-uncontrollability may be defined similarly as

$$J_{\overline{XZ^*}}: \ (\exists (X, Z^*): \forall E, \ \bar{\exists} U: Z(X, U, E) = Z^*); \tag{2.3.16}$$

The Z*E-uncontrollability:

$$J_{\overline{Z^*E}}: \ (\exists (Z^*, E): \forall X, \ \bar{\exists} U: Z(X, U, E) = Z^*). \tag{2.3.17}$$

And, finally, the XEZ*-uncontrollability:

$$J_{\overline{XEZ^*}}: \ (\exists (X, E, Z^*): \bar{\exists} U: Z(X, U, E) = Z^*). \tag{2.3.18}$$

In Fig. 2.3.2 the zones of uncontrollability are shown crosshatched. Evidently, they also include the zones shown in Fig. 2.3.1. However, one should not mix up the areas of the uncontrollability of this kind with intersection of the areas of X-, E-, and Z*-uncontrollability shown in Figs. 2.3.1*b*, *c*.

If it might be possible with the aid of experts to estimate the occurrence (probability of emergence) of all the mentioned kinds of uncontrollability, then calculation of the object controllability could be reduced to the formula

$$P = 1 - \sum_{i=1}^{7} p_i, \tag{2.3.19}$$

where p_i is the probability of emergence of the i-th kind of uncontrollability (they are seven altogether). This would be the algorithm φ in (2.3.9). Disposing of such an algorithm, one can select the most controllable object out of a certain finite set of objects-contenders.

So let us turn again to the expert method: it is required to estimate the values p_i ($i = 1, \ldots, 7$) included into the formula of controllability (2.3.19).

2.4. The Expert Method for Evaluation of the Object's Controllability

Thus, it is necessary to formulate the questions put before the selected N experts in such a way that their answers would allow taking the most reliable decision on the object controllability within the given limits. First of all, this requires to produce a finite number of alternatives, versions of the object $\mathbf{F}_1^o, \ldots, \mathbf{F}_q^o$, i.e. the descriptions

$$\mathbf{Y} = \mathbf{F}_\nu^o(\mathbf{X}_\nu, \mathbf{U}_\nu, \mathbf{E}_\nu), \quad \nu = 1, \ldots, q, \tag{2.4.1}$$

where $\mathbf{X}_\nu$, $\mathbf{U}_\nu$, and $\mathbf{E}_\nu$ are respectively the observed, controlled, and unobserved inputs of the ν-th version of the object. Versions (2.4.1) are produced on the basis of preliminary analysis of the medium and the objectives.

Therefore, the problem of object determination is reduced to selecting one of the alternatives (2.4.1). The initial data for the selection are contained in the experts' answers to the questions that help determine the controllability of each version of the object (2.4.1). The questions to the experts should be formulated so that the answers could allow establishing the probability of each of the types of uncontrollability. The situations S in which the objectives are unachievable, i.e. the object is uncontrollable, are to be revealed.

The expertise proceeds through several stages. First, the experts identify and describe the uncontrollable situations. Then a full list of such situations categorized by the types of uncontrollability is compiled. The list is submitted to the experts to obtain their estimates of the probability of each of the situations. If the estimates are not too different, they are taken for calculating controllability of the object. But when the discrepancy is large, the experts are given a possibility to change their estimates on the moot points considering those of the others. This is repeated until the estimates are no longer much diverse. Then the controllability is calculated according to the final estimates.

Now let us discuss each stage in particular. At the first one, every expert is expected to describe all the uncontrollable situations within each type of uncontrollability:

$$J_{\overline{X}}: (S^1{}_1, \ldots, S^1{}_{l_1}), \ldots, J_{\overline{XEZ}^*}: (S^7{}_1, \ldots, S^7{}_{l_7}), \quad (2.4.2)$$

where S^i_j is the j-th uncontrollable situation of the i-th type of uncontrollability.

Apparently, each expert will have his own set of such situations. The situations are merged into a common list of uncontrollable situations. It will have the same form as (2.4.2), but with different numbers $l_1, \ldots, l_7$.

When compiling the list, the situations should be checked for not being included more than once. This is necessary because, for instance, the situations from $J_{\overline{XE}}$ may prove to be mixed up with the situations simultaneously belonging to $J_{\overline{X}}$ and $J_{\overline{E}}$ (see the intersections in Figs. 2.3.1*b*, *c* and 2.3.2*a*, *b*).

Now each expert is offered the full list of all uncontrollable situations

$$S_1, \ldots, S_n \tag{2.4.3}$$

with no account of their belonging to the types of uncontrollability (this only served the purpose of identification of the situations (2.4.3) by the experts). Each j-th expert attributes to each situation S_i a figure

$$0 \leqslant g^j{}_i, \quad i = 1, \ldots, n; \quad j = 1, \ldots, N \tag{2.4.4}$$

being, in his opinion, the probability of emergence of this particular situation in the process of the object control. There is an obvious constraint

$$\sum_{i=1}^{n} g^j{}_i \leqslant 1, \quad j = 1, \ldots, N.$$

Let us note that the sum

$$\overline{P}_j = \sum_{i=1}^{n} g^j{}_i, \quad j = 1, \ldots, N \tag{2.4.5}$$

is the opinion of the j-th expert on the uncontrollability of the object.

But before calculating the mean value of the object's uncontrollability, the concordance of the experts' estimates should be checked up. This may be checked up by the dispersions of individual estimates of the situations

$$\hat{D}_i = \frac{1}{N-1} \sum_{j=1}^{n} \left(g^j{}_i - \frac{1}{N} \sum_{l=1}^{N} g^j{}_i \right)^2, \quad i = 1, \ldots, n. \tag{2.4.6}$$

In case there are such situations S_i for which the opinions of the experts on the probability of their emergence diverge greatly, i.e.

$$\hat{D}_i > \delta, \quad i = 1, \ldots, n, \tag{2.4.7}$$

where $\delta > 0$ is a certain given threshold, then the experts are asked either to reconsider or confirm their estimates with due regard for the opinions of the other experts (the

list of such situations and the opinions of the others are presented to each expert). The procedure is repeated until

$$\hat{D}_i \leqslant \delta, \quad i = 1, \ldots, n.$$

Now the combined experts' judgement on the uncontrollability of the object may be viewed as

$$\overline{P} = \left(\sum_{j=1}^{N} k_j \overline{P}_j \right) \Big/ \sum_{j=1}^{N} k_j, \tag{2.4.8}$$

where k_j is the competence of the j-th expert (2.2.4) and $\overline{P}_j$ is the estimate of the object's uncontrollability by the j-th expert (2.4.5).

This is the final stage of the expert evaluation of the given version of the objects.

Having obtained such experts' estimates on all the q versions (2.4.1) of the object, we can find the values of their uncontrollability $\overline{P}^1, \ldots, \overline{P}^q$. Obviously, the best version of the object has the minimal uncontrollability. If

$$\overline{P}^l = \min_{\nu=1, \ldots, q} \overline{P}^\nu,$$

then the l-th version of the object should be considered to be optimal, i.e. $\mathbf{F}^0 = \mathbf{F}_l^q$.

Naturally, the dispersion of probabilities $\overline{P}^\nu$ ($\nu = 1, \ldots, q$) should be taken into account. In case several versions of the object prove to be statistically identical, then either the number of experts should be increased, or they should be provided with extra data on the medium, objectives and resources, or new criteria should be added for the estimations allowing unambiguous selection of the object version.

Comments on References to Chapter 2

The primacy of the objective over the object is due above all to the fact that any object is a system which according to [1, 2] is a set of elements with a certain relationship. This relationship has a common feature determined by a set of the objectives to be realized in the object. Therefore, any object of control is different from the system as a whole because the object consists of elements which are connected to each other by a relation that enables the

object to exist in states which come from a set of objectives. This restriction on the states of the object to correspond to the set of objectives makes the object depend on the control objectives.

Clearly, the expert evaluation method discussed in § 2.2 allows solving problems formally unsolvable at least because they cannot be formally presented. And there are many more of them today than the proponents of the formal approach would like. In fact, more and more commonly now, we have to control objects long before we can study them in full, and have to switch over to control other objects without reaching enough knowledge on the ones left behind. It is natural in the circumstances that only the expert method allows solving the set problems. The 'price' is often a greater dispersion of estimates which should not be, however, confusing since there are effective methods of stochastic decision-making (see, for instance, quite accessible books on the theory of statistic decisions [3] or [4]).

Therefore, the use of the expert solutions is not at all 'sweeping the dust under the rug' as it may seem. Combined with the methods of the theory of statistic decisions, the expert method brings quite acceptable results.

There are a number of useful textbooks on the method of expert estimates. Let us note [5] and [6] among them. An easy and comprehensive book [7] contains chapters recommended for the initial study.

The concept of object controllability introduced in § 2.3 differs from a familiar concept of controllability of the dynamic system (e.g. see [8]) which implies the existence of control taking the object into a required state. The concept proposed here obviously generalizes the known one including the stochastic case. Therefore, it seems to be more correct to term the introduced characteristic of the object as statistic controllability, which would make, however, the terms too cumbersome.

References

1. Uemov, A.I. *Things, Properties, and Relationships.* USSR Acad. Sc. Publ., Moscow, 1963 (in Russian).
2. Uemov, A. I. Systems and systems parameters. In: *Problems of Formal Analysis of Systems.* Visshaya shkola, Moscow, 1968 (in Russian).
3. Chernoff, H., Moses, L.E. *Elementary Decision Theory.* Wiley, Inc., New York, 1959.
4. Groot, M.H. de. *Optimal Statistical Decisions.* McGraw-Hill Book Co., New York, 1970.
5. Beshelev, S.D., Gurovich, F.G. *Mathematical and Statistical Methods of Expert Evaluation.* Statistika, 1974 (in Russian).
6. Expert evaluations in technological prognostication. Naukova dumka, Kiev, 1974 (in Russian).
7. Shlyapentokh, V. *How Today They Study Tomorrow.* Sov. Rossiya, Moscow, 1975.
8. Meditch, J.S. *Stochastic Optimal Linear Estimation and Control.* McGraw-Hill Book Co., New York, 1969.

Chapter 3 Structural Synthesis of the Object's Model

The structure of the model of the object of control determines the form and nature of connection between the inputs (X and U) and outputs (Y) of the object's model with no regard to concrete values of the parameters to be considered at the later stages of control, the stages of identification (see Chapter 4) and experiments design (see Chapter 5). We shall divide the process of structural synthesis of the object's model into the following stages:

—determination of the object's inputs and outputs;
—expert ranking of the inputs and outputs;
—decomposition of the model;
—selection of the model's structural elements.

Let us consider each stage in particular.

3.1. Determination of the Object's Inputs and Outputs

In order to determine the structure of the model as a multi-port (i.e. a converter of the 'black box' type with many inputs and outputs), it is necessary to recognize which inputs and outputs of the object are going to be included into its model.

First of all, this requires to find out all possible 'contenders' to be inputs and outputs, and to select the most essential of them that will form the multi-port.

But here an immediate question arises: what factors should be considered essential? The answer is unambiguous: since the model of the object is produced for the purposes of control, the essential ones are the factors that bear the greatest effect on the realization of the objectives of control in the object. This is due to that in the determination of the structure of a control system first of all

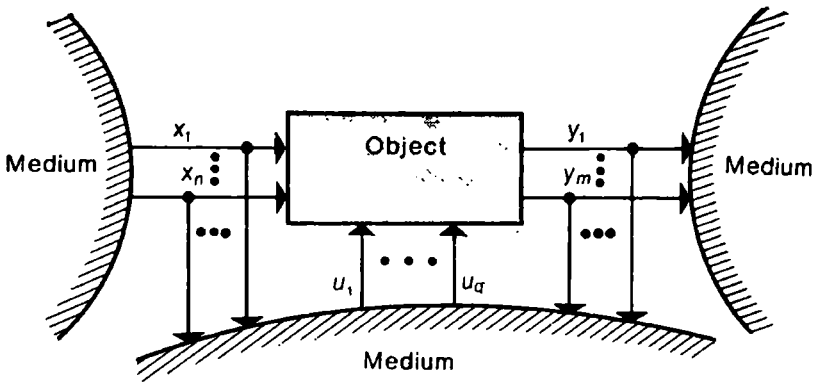

Fig. 3.1.1. Interaction between the object ('black box') and the medium

it is necessary to know which effects the object of control will (and may) be subjected to and how the results of those effects are connected with the achievement of the objectives of the object control.

Generally, the object is connected with the medium by an infinite number of connections (Fig. 3.1.1) influencing its state. The strongest, most essential ties should be identified at the stage of the model's synthesis so that turning down the weaker effects we could operate with a minimal number of the object-medium interactions. However, before separating them, we have to list all the connections. Obviously, those are going to be only the observed ties, since no unobserved ones can be included into the model. Let us group them into three types:

—uncontrollable but observed connections (X) characterizing the effect of the medium on the object (the uncontrollable inputs are among these);

—controlled connections (U) with the aid of which the state of the object may be purposefully changed (the controlled inputs are among these);

—informing connections (Y) helping to check the state of the object (its outputs are among these).

Let us consider each type of connections separately.

The uncontrollable connections X should meet the following requirements.

First, they should have some influence on the realization of the objectives $\{Z^*\}$ in the controlled object. We

can make a mental experiment to see the influence. Let the object be in a state $\mathbf{Y}^*$ corresponding to a given objective $\mathbf{Z}^*$, i.e. the objective has been achieved (it doesn't matter how, e.g. by means of an appropriate control $\mathbf{U}^*$). Now let a parameter x_i (characterizing the state of an i-th object-medium connection) change, i.e. let there be a change in the medium (but only in its i-th connection with the object). If that change of the medium can change the state of the object so that the objective $\mathbf{Z}^*$ cannot be reached, then the connection x_i is essential for the object control since it changes the process of realization of the objectives in the object. Otherwise, this connection may be neglected.

Second, such connections should be measurable efficiently and easily. If the parameter x_i cannot be measured, i.e. there are no methods to measure it or they are too involved or costly, then this connection cannot be taken into account in the control synthesis and it should be regarded as a random disturbance ε.

Thus when choosing the factors of the medium it is necessary to consider those factors for the control which are essential for the realization of the control objective as well as factors which can be effectively measured and are verifiable.

Since there are no models of the medium and the object, the selection of factors of the medium to be taken into account is to be performed by an expert method, i.e. with the help of specialists (experts) knowing the object and the medium well. The level of the experts' knowledge is crucial for the efficiency of the process of finding the essential connections of the object with the medium.

Suppose according to that method n connections have been selected:

$$x_1, \ldots, x_n. \tag{3.1.1}$$

This means that the content of each connection has been found out (e.g. x_1 is the temperature of the medium, x_2 is the atmospheric pressure, x_3 is the voltage, etc.) and it has been established that these connections influence in a way the realization of the objectives in the object.

Now let us consider the process of selecting the controlled connections $\mathbf{U}$, the channels of the projected con-

trol. The most important criteria for the selection are:

—the effect of the connection on the state of the object, i.e. a possibility to compensate for some negative changes in that state caused by various factors of both the medium **X** and the object $\mathbf{F}^{o}$;

—controllability, i.e. the possibility of changing efficiently and without undue expenditure the state **U** of the channels controlling the object.

These are the criteria for the determination of the content of connections

$$u_1, \ldots, u_q \tag{3.1.2}$$

by the expert method also (e.g. u_1 is fuel consumption, u_2 is the instruction to switch on some aggregate, etc.).

And finally, the informing connections **Y**. They should meet two natural requirements:

—to bear the data on the realization (or non-realization) of the objectives in the object of control;

—to be measurable (checked) reliably and efficiently.

The thus selected informing connections

$$y_1, \ldots, y_m \tag{3.1.3}$$

(e.g. y_1 is the temperature in the object, y_2 is its speed, etc.) complete the description of the object as a 'black box', i.e. as a multi-port with an unknown (so far) internal structure.

Therefore, the stage of determination of the object's inputs and outputs consists in describing the content of all the observed inputs and outputs of the object, regardless of its internal structure, which are connected with the realization of the objectives in the object. Apparently, their number should be larger than that realized in the model of the object. Such redundancy is required for the efficient selection of the most essential factors (connections) at the next stage.

3.2. Expert Ranking of the Inputs and Outputs

The connections determined in section 3.1 are not yet the inputs and outputs of the model. These are only the 'contenders'. They are going to become inputs and out-

puts after the most essential ones will be isolated and inessential discarded. As a matter of fact, the model should not reflect absolutely all the features of the object, i.e. it should not (and cannot) be entirely adequate to the object. For the purposes of control, it is quite sufficient to determine the model's most essential connections with the medium, those which are necessary for control, i.e. the model may be an approximate one. That is why the process of selecting the most essential factors is preceded by the procedure of their ranking according to their significance for the realization of the objectives of control.

Let us consider the process of ranking for the factors x_i ($i = 1, \ldots, n$). The factors **U** and **Y** are to be ranked in the same way.

First, let us define the concept of ranking. We shall assume by 'ranking' a procedure of putting the factors x_i in the order of their significance: the most essential one should be the first, the following one should be less essential, but the most important among the rest, etc. The ranked sequence thus produced has the form

$$x_{i_1}, x_{i_2}, \ldots, x_{i_n}, \tag{3.2.1}$$

where i_1 is the number of the most essential factor, i_2 is the number of the less essential one, etc., up to the number i_n of the least essential factor in this sequence.

However, it is possible, and sometimes more convenient, to try another approach. Every factor x_i may be assigned an integer, its rank k_i, i.e. the number of the factor in the ranked sequence (3.2.1):

$$\begin{matrix} x_1, x_2, \ldots, x_n \\ k_1, k_2, \ldots, k_n. \end{matrix} \tag{3.2.2}$$

Obviously, the first rank ($k_i = 1$) is to be assigned to the input x_{i_1} that is the most important for the realization of the objective in the object. The second and the following ranks ($k_i = 2$, etc.) in the order of their diminishing importance are to be assigned to the less essential inputs. Therefore, the connection of ranked sequence (3.2.1) and ranks (3.2.2) may be defined by a simple and evident

formula

$$k_{i_j} = j \quad (i, j = 1, \ldots, n), \tag{3.2.3}$$

i.e. the rank of the factor with the number i_j is equal to j.

For instance, if ranks (3.2.2) turned out to be

$$\begin{aligned} x_i &= x_1, \ x_2, \ x_3, \ x_4, \ x_5, \\ k_i &= 3, \ \ 1, \ \ 5, \ \ 4, \ \ 2, \end{aligned} \tag{3.2.2a}$$

then the ranked sequence is x_2, x_5, x_1, x_4, x_3. Indeed, it follows from (3.2.2a) that the second factor has the first rank ($k_i = 1$), the fifth has the second, etc. Now in case we are to design a control system with a limited number of inputs from the medium (e.g. $n = 3$), the essential factors are selected from (3.2.2a) in an obvious sequence. These are x_2, x_5 and x_1. The fourth and the fifth factors are to be neglected, but it is evident that the adverse effect of such a decision is going to be minimal since those are the least essential factors.

The problem of producing ranked sequence (3.2.1) or the equivalent problem of determining ranks (3.2.2) may be solved by experts. Thus, it is reduced to the organization of the experts' opinion poll and processing the results in order to obtain the sought ranks and estimate their reliability, i.e. the agreement of the experts' opinions.

We shall consider two methods of expert ranking: the direct ranking method and the method of pair comparisons. According to the first method, the experts assign ranks directly to the factors put for ranking, while the pair comparisons of factors are used in the second method which simplifies the job for the experts, but requires further processing of the results to obtain the ranked sequence.

The method of direct ranking. Let N experts rank n factors $x_1, \ldots, x_n$. Each expert assigns a rank, an integer from 1 to n, to each factor. Thus, a j-th expert (E_j) assigns to an i-th factor (x_i) a rank k_{ij}. The result is the experts' opinion matrix $N \times n$

$$\begin{array}{c|cccc} & x_1 & x_2 & \ldots & x_n \\ E_1 & k_{11} & k_{21} & \ldots & k_{n1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ E_N & k_{1N} & k_{2N} & \ldots & k_{nN} \end{array}, \tag{3.2.4}$$

where the numbers of rows correspond to those of experts, and the numbers of columns to those of the factors to be ranked, i.e. the j-th row is the opinion of the j-th expert E_j, and the i-th column combines the opinions of all the experts concerning the i-th factor (x_i).

The following requirements should be met by the experts while assigning the ranks:

—first, the sum of the ranks assigned to all the factors by each expert should be the same. It implies that the sum of the elements for any row of the matrix should be

$$\sum_{i=1}^{n} k_{ij} = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \quad (j = 1, \ldots, N);$$

—second, in case an expert considers some q factors to be equivalent or of the same importance, they should be assigned the same rank. This rank is equal to the mean value of q integer ranks which would appear provided the expert managed ranking the factors in a sequence. For instance, the equivalence of four factors $(q = 4)$: x_1, x_2, x_5, x_6, if they occupy the fifth place in the ranked sequence, results in equal ranks

$$k_1 = k_2 = k_5 = k_6 = (5 + 6 + 7 + 8)/4 = 6.5.$$

Clearly, this is the case when the ranks may be fractional.

Now to finally determine the sought ranks, it is necessary to compute the mean ranks for each factor:

$$\bar{k}_i = \frac{1}{N} \sum_{j=1}^{N} k_{ij}.$$

These ranks allow putting the factors into the proper ranked sequence. The first place is to be occupied there by the factor with the minimal mean rank,

$$\bar{k}_l = \min_{i=1, \ldots, n} \{\bar{k}_i\},$$

i.e. the factor x_l, the second place belongs to the factor with the minimal mean rank among the rest, etc. The resulting ranks produce the ranked factor sequence reflecting the 'averaged' opinion of the group of N experts.

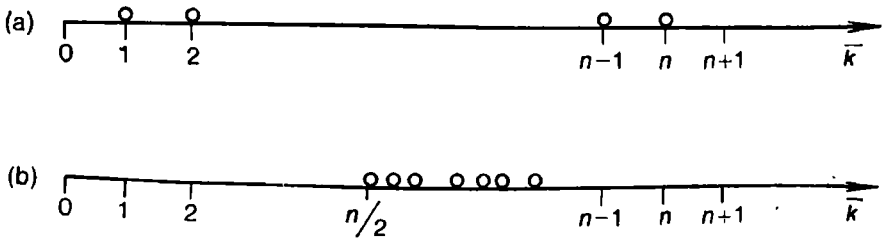

Fig. 3.2.1. Mean ranks on the number-scale axis: when the opinions of the experts agree completely (*a*); when they agree partially (*b*)

Obviously, far from every outcome of the experts' opinion poll should be regarded as satisfactory. Thus, in case there is a wide gap in the experts' opinions (e.g. half of the experts assigned to the same factor x_i the highest rank, while the other half the lowest one), then such ranking cannot constitute the basis for decision-making in the selection of the essential factors. Therefore, in order to evaluate the results of any experts' opinion poll, a certain criterion is to be introduced characterizing the agreement of the experts. The higher the degree of agreement, the more we can believe the results of the poll, and vice versa.

It is convenient to determine the agreement of the experts as a degree of dissipation of mean ranks $\bar{k}_i$ ($i = 1, \ldots, n$). Clearly, if the opinions of the experts are completely the same, the mean ranks are unequal integers (the case of equal ranks is not considered here yet). It is shown diagrammatically in Fig. 3.2.1*a*, where the small circles present the location of mean ranks on the number-scale axis. But if the experts' opinions are completely diverse, the mean ranks are approximately equal to $n/2$. When the opinions agree partially, the ranks tend to group around $n/2$ (see Fig. 3.2.1*b*).

The degree of dissipation is conveniently determined by mathematical dispersion. By definition, the dispersion of the mean ranks is

$$D(\bar{k}) = \frac{1}{n} \sum_{i=1}^{n} (\bar{k}_i - M(\bar{k})),$$

where

$$M(\bar{k}) = \frac{1}{n} \sum_{i=1}^{n} k_i = \frac{n+1}{2}$$

is the mathematical expectation of the mean rank.

Let us determine the maximal dispersion (it is easy to see that it occurs when the experts' opinions are completely the same):

$$D_{\max} = \frac{1}{n} \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right)^2 = \frac{n^2-1}{12}.$$

The criterion of the experts' agreement may be conveniently presented as a ratio

$$W = \frac{D(\bar{k})}{D_{\max}} = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} \left(\bar{k}_i - \frac{n+1}{2} \right)^2.$$

Clearly, $0 \leqslant W \leqslant 1$. When $W = 0$, the experts disagree completely, but when $W = 1$, they agree completely. Consequently, the value of W characterizes the degree of the experts' agreement. The closer is W to unity, i.e. the more unanimous are the experts, the more reliable is the outcome of the ranking.

Let us note that the experts should offer their opinion independently from each other, i.e. they should not know the opinions of the others before ranking. Otherwise, there is a possibility of having dependent opinions, though rising the degree of agreement W, but not improving the results of the poll.

To get a glimpse of what a concrete value of the agreement criterion might be, which is never equal to zero or to unity, and to have an idea of whether it is great or low, let us try the following simple method. Suppose m out of the N experts are absolutely competent, and the rest $(N - m)$ are absolutely incompetent, i.e. they make their decisions randomly (though all that, naturally, cannot happen). Now the dispersion of the mean ranks

$$D(\bar{k}) = \frac{1}{N} [m D_{\max} + (N - m) 0] = \frac{m}{N} D_{\max}.$$

Dividing the result by D_{max} we shall get $W = m/N$. Therefore, the value of W depends on the number of completely competent experts (if such were existent). Thus, when $W = 0.3$, it is possible to believe that 30% of the experts were absolutely competent, and the rest 70% took their decisions randomly, which might (and might not!) be fateful for the ultimate ranked sequence.

Let us illustrate the method applying it to a joking example of ranking girls. Let us consider a common thing, a talk of three boys about their girl-friends. Naturally, each of them prefers his own. Is it possible to apply the method of expert evaluation in order to rank the girls objectively according to some criterion (if such ranking can really happen)? It proves that it is, in spite of the obvious bias of the experts, our boy-friends. Let x_i be an i-th girl ($i = 1, 2, 3$), and E_j be a j-th expert ($j = 1, 2, 3$). Let there also be an a priori true ranked sequence

$$x_1 \succ x_2 \succ x_3.$$

Here the preference symbol $\succ$ means 'cleverer than' or 'better home-maker than' or 'nicer than', etc., depending on the criterion picked by the boys. But only E_1 agrees with the above sequence. E_2 believes that

$$x_2 \succ x_1 \succ x_3,$$

and E_3 has an opinion of his own:

$$x_3 \succ x_1 \succ x_2.$$

Obviously, each of the boys agrees with the true sequence in the part that does not concern his respective girl-friend. The opinions may be presented together in a matrix of the (3.2.4) type:

$$\begin{array}{c|ccc} & x_1 & x_2 & x_3 \\ E_1 & 1 & 2 & 3 \\ E_2 & 2 & 1 & 3 \\ E_3 & 2 & 3 & 1 \end{array}.$$

It is easy to see that the mean ranks of the girls, $\bar{k}_1 = 5/3$, $\bar{k}_2 = 2$, $\bar{k}_3 = 7/3$ correspond to the true ranking, i.e. the experts' opinion poll produced a correct answer.

Now the agreement of our experts

$$W = D(\bar{k})/D_{max} = (2/27)/(2/3) = 1/9,$$

i.e. their opinions agree very poorly, as it could have been expected. Nonetheless, the ultimate ranking proved to be correct. That occurred due to the averaging of the experts' opinions which smoothed out their individual inclinations and thus excluded their errors.

This result shows the efficiency of the method of experts' evaluation even in case the experts clearly disagree. It is important that the disagreement should concern different ranked factors.

The lack in the experts' agreement may be an evidence, on the one hand, of the experts' incompetence due to the novelty of the object or poor research on it. On the other hand, it may result from the complexity of the object, making it difficult to rank the factors. It is easier for the experts to compare the importance of factors in pairs, i.e. to indicate the rank of which factor is going to be greater. These are the situations when the method of pair comparisons is used.

The method of pair comparisons. The expert has to rank the factors in pairs, i.e. to assign to each pair of factors x_i and x_l a number

$$q_{il} = \begin{cases} 1, & \text{if } x_i \succ x_l, \\ 0, & \text{if } x_i \sim x_l, \\ -1, & \text{if } x_i \prec x_l. \end{cases}$$

The expression $x_i \succ x_l$ means that the i-th factor is more preferable in ranking than the l-th. The symbol $\sim$ is the sign of equivalence of factors from the point of view of ranking (or the sign of refusal to rank them). Obviously, the numbers $q_{ll} = -q_{li}$.

Thus, each j-th expert presents his opinion in the form of a matrix

$$Q^j = \| q^j_{il} \|, \quad (i, l = 1, \ldots, n; \; j = 1, \ldots, N),$$

where the upper index is the number of the expert (see Fig. 3.2.2).

Let us try and average the experts' opinions. It is sufficient for that to produce an averaged matrix $n \times n$

$$\overline{Q} = \| \overline{q}_{il} \|,$$

where

$$\overline{q}_{il} = \frac{1}{N} \sum_{j=1}^{N} q^j_{il}$$

is the mean preference of the i-th factor to the l-th one. That represents the opinion of the given group of experts.

	x_1	• • •	x_n
x_1	0	• • •	q^j_{1n}
•	•		•
•	•		•
•	•		•
x_n	q^j_{n1}	• • •	0

Fig. 3.2.2. Preference table for the j-th expert by the method of pair comparisons

	x_1	x_2	x_3
x_1	0	1	−1
x_2	−1	0	1
x_3	1	−1	0

Fig. 3.2.3. An example of a contradictive matrix

Let us determine the experts' agreement. Similarly to the first method, it is natural to select the dispersion of the values $\bar{q}_{il}$ as a criterion of agreement. Since their mean value is equal to zero, we obtain

$$D(\bar{q}) = \frac{1}{n(n-1)} \sum_{i,\, l=1}^{n} (\bar{q}_{il})^2,$$

where the sum covers the whole matrix $\bar{Q}$. The maximal value of dispersion $D_{\max} = 1$ occurs when the experts agree completely. Now we can introduce another criterion of agreement, the ratio of dispersion of the mean values to maximal dispersion

$$W = \frac{1}{n(n-1)} \sum_{i,\, l=1}^{n} (\bar{q}_{il})^2.$$

Obviously, $0 \leqslant W \leqslant 1$; when $W = 1$, the experts' opinions are in complete agreement, and when $W = 0$, they contradict each other.

However, there may be discrepancies even when the experts completely agree. An example of a matrix of the kind is shown in Fig. 3.2.3, where the contradiction is $x_1 \succ x_2 \succ x_3 \succ x_1$, i.e. it turned out that $x_1 \succ x_3$ and $x_3 \succ x_1$ simultaneously. It is absolutely necessary to reveal such contradictions both in the averaged matrix and in the matrices of each of the experts. This is carried

out on the basis of the following rather obvious rule which is called the *rule of transitivity*. For preferences, it has the form

$$\begin{aligned} &\text{if } x_1 \succ x_2 \text{ and } x_2 \succ x_3, \text{ then } x_1 \succ x_3; \\ \text{and} \quad &\text{if } x_1 \sim x_2 \text{ and } x_2 \sim x_3, \text{ then } x_1 \sim x_3 \end{aligned} \tag{3.2.5}$$

for equivalences.

The matrices presenting the opinion of each expert should meet this transitivity rule. In case the contradictions are revealed, the matrices are returned to the experts for correction.

To determine the sought ranks of the ranked factors, we should have a rule to compute the ranks using the matrix $\bar{Q}$. There may be many such rules. Let us consider some of them.

Rule 1. It is clear that the value $\bar{q}_{il}$ shows the degree of preference of the i-th factor to the l-th one. Let us determine the mean preference of each factor to all the others:

$$\bar{q}_i = \frac{1}{n} \sum_{l=1}^{n} \bar{q}_{il} \quad (i = 1, \ldots, n). \tag{3.2.6}$$

It is natural to assume that the factor with the maximal mean preference has to have the first rank. Thus, when

$$\bar{q}_v = \max_{i=1, \ldots, n} \{\bar{q}_i\}$$

the first rank is assigned to the factor x_v, i.e. $k_v = 1$. The ranks of the other factors are produced similarly.

This rule, however, smoothes out the preferences too much. Thus, a factor with a number of obvious (i.e. large) preferences easily disclosed by the experts will get the first rank only because its secondary quality with regard to other factors was not seen clearly. This is the case when another rule is often applied.

Rule 2. This rule is based on the idea of accentuating contrasts. A threshold δ is introduced for the purpose. In case a preference is higher than this threshold, it is evident, and when it is lower, it is questionable, i.e. the factors are rather equivalent. Therefore, the following is applied to transform the matrix of the mean preferences $\bar{Q}$

	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
x_1	0	-0.7	-0.44	-0.52	0	-1	0	0	0	-1	0	-1	0	-1	0	-1
x_2	0.7	0	-0.74	0.5	1	0	-1	0	1	0	-1	1	1	0	-1	0
x_3	0.44	0.74	0	-0.78	0	1	0	-1	0	1	0	-1	0	1	0	-1
x_4	0.52	-0.5	0.78	0	0	0	1	0	1	-1	1	0	1	0	1	0
			(a)				(b)				(c)				(d)	

Fig. 3.2.4. A matrix of mean preferences (*a*); contrast matrices for $\delta = 0.7$ (*b*), $\delta = 0.5$ (*c*), and $\delta^* = 0.52$ (*d*)

into a contrast matrix, the elements of which are

$$\varphi_{il} = \varphi(q_{il}), \quad (i \neq l = 1, \ldots, n),$$

where

$$\varphi(q) = \begin{cases} -1, & \text{if } q \leqslant -\delta, \\ 0, & \text{if } |q| < \delta, \\ 1, & \text{if } q \geqslant \delta. \end{cases}$$

The transformation completely depends on the threshold δ ($0 < \delta \leqslant 1$). When $\delta = 1$, the contrast matrix usually becomes all zeroes and all the factors are assumed to be equivalent. When $\delta = 0$, it is filled with unities all over, but contradictions appear almost unavoidably in the form of the discrepancies in preference transitivities (3.2.5). Therefore, in selecting the threshold δ, we should have it in mind that its increase results in non-ranking (all factors become equivalent), and its decrease leads to the danger of contradictions. One of the possible recommendations to determine the optimal threshold is to select the value of δ at the 'threshold of contradictions', i.e. a minimal value δ^* such that its small decrease brings about contradictions.

Example. Let a mean preference matrix have the form shown in Fig. 3.2.4*a*. If $\delta = 0.7$, the contrast matrix (*b*) produces the following ranked sequence:

$$x_4 \succ x_3 \succ x_2 \succ x_1. \qquad (*)$$

A zero in this matrix signifies not only equivalence, but also a mild preference. Therefore, the obtained ranked sequence is not contradictive. Let us lower the threshold in compliance with

the algorithm. When $\delta = 0.5$, the contrast matrix (*c*) already contains a contradiction since $x_4 \succ x_3 \succ x_2 \succ x_4$. The minimal threshold, at which there is no contradiction, is $\delta^* = 0.52$ (see Fig. 3.2.4*d*), which produces the ranked sequence (*), i.e. at the thresholds $\delta = 0.52$ or 0.7 we have the same ranked sequence (*).

Let us apply the first rule to this example. The matrix in Fig. 3.2.4*a* by means of (3.2.6) produces $\bar{q}_1 = -1.66$; $\bar{q}_2 = 0.46$; $\bar{q}_3 = 0.40$; $\bar{q}_4 = 0.8$. Hence, according to the first rule, we obtain a ranked sequence

$$x_4 \succ x_2 \succ x_3 \succ x_1. \qquad (**)$$

The result clearly differs from (*). This means that either $x_3 \sim x_2$ or a more accurate inference requires new data to elucidate the relationship between x_2 and x_3, i.e. to decide which of the two ranked sequences (*) or (**) is true.

3.3. Decomposition of the Model

So far we have considered the object as a 'black box' with nothing known about its contents, and whose model was presented as a multi-port with an unknown structure.

At the next stage of the structural synthesis of the model, it is natural to use some additional data on the object as a system of interacting elements. For instance, such an object of control as a factory is a system of interacting workshops, a technological process is a system of interacting operations, an organism is a system of interacting organs, a collective is a system of its interacting members, etc.

The data on such structural features of the object are the initial ones to decompose the model of the object, i.e. to 'split' the model into its interacting elements which represent the complex multi-element structure of the object. An example of such a decomposition is shown in Fig. 3.3.1 where a model with three inputs and three outputs contains five simpler elements (*1-5*).

Using a priori data on the structure of the object, we may simplify the synthesis of the model. This is what decomposition is for. The simplification may go along two routes. On the one hand, each element of the decomposed system should be simpler than that of the initial (non-decomposed) system, which makes easier the syn-

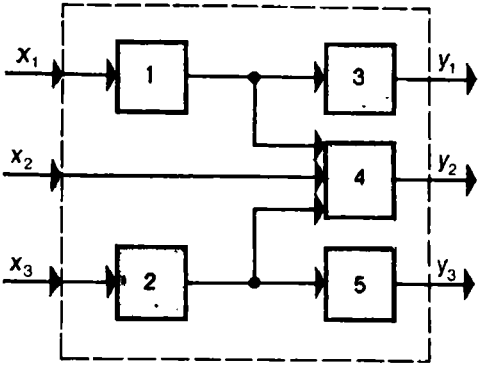

Fig. 3.3.1. An example of model decomposition

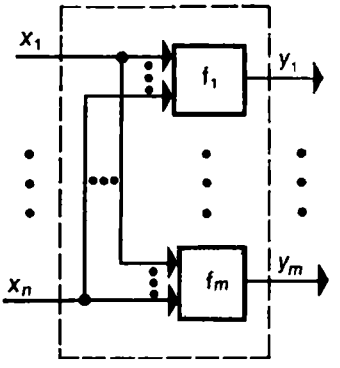

Fig. 3.3.2. Decomposition of $(n \times m)$-port without a priori data on the structure of the object

thesis of the elements and of the whole system in the long run. On the other hand, the problem of the synthesis may be simplified despite the complexity of the elements, if the number of inputs and outputs of each element is made lower compared with the initial object.

Let us make clear the latter. Let us introduce a numerical measure of complexity of the model as of a 'black box'. It is natural to base this measure on labour consumption of the model synthesis, i.e. on the expenditures necessary to produce the model. Let us call this measure the 'complexity'. At the stage of the analysis (the 'black box' stage), the complexity should take into account only the number n of inputs and the number m of outputs of the model (for simplicity, we shall not divide the inputs into controllable and uncontrollable). So let it have the form

$$L = L(n, m) \tag{3.3.1}$$

We may put forward the following requirements for this function. First, it should be monotonically increasing with respect to both n and m. Second, the number of the inputs n should have a greater effect on the complexity than the number of the outputs. And the last: the function should be assumed to be additive, i.e. if the object consists of several (g) subsystems, its complexity equals

the sum of complexities of these subsystems, i.e.

$$L = \sum_{i=1}^{g} L_i, \tag{3.3.2}$$

where L_i is the complexity of the i-th subsystem of the object. (Let us note that when g is great, this assumption, strictly speaking, is wrong, since a complex object is not a sum of its parts. Expression (3.3.2) does not take into account the difficulties of the process of subdividing the object into its components.) All these considerations are satisfied, for instance, by the following function (though many others may be presented as well):

$$L = n^{\gamma}m, \tag{3.3.3}$$

where $\gamma > 1$, since the number of inputs has a greater effect on complexity than the number of outputs. The value of γ should be determined according to how many times the labour consumption of the model synthesis increases with an increase in the number of its inputs by unity. Let the labour consumption be increased p times, then it follows from (3.3.3) that

$$(n+1)^{\gamma}\, m = pn^{\gamma}m,$$

from which it is easy to obtain

$$\gamma = \frac{\log p}{\log (n+1/n)}. \tag{3.3.4}$$

Applying (3.3.3) to the example of the model decomposition shown in Fig. 3.3.1, we shall have for the whole system

$$L = 3^{\gamma} \cdot 3 = 3^{\gamma+1},$$

and for the decomposed system

$$L_d = \sum_{i=1}^{5} L_i = 1+1+1+3^{\gamma}+1 = 4+3^{\gamma}.$$

It is clear that $L > L_d$ even at $\gamma = 1$, when $L = 9$ and $L_d = 7$. In case γ is unknown, it may be assumed equal to unity, which is obviously understating the complexity L.

Now the process of model decomposition may be considered as a process of minimizing its complexity, i.e. as

the solution of the following minimization problem:

$$L_d \to \min_{D \in \{D\}} \Rightarrow D^*, \tag{3.3.5}$$

where D is the operation of decomposition, $\{D\}$ is the set of decompositions admissible for the given object; D^* is the optimal decomposition minimizing the complexity L of the decomposed system.

Let us apply formula (3.3.3) to the decomposition without any a priori knowledge on the structure of the object (as shown in Fig. 3.3.2). We shall have

$$L_d = \sum_{i=1}^{m} L_i = \sum_{i=1}^{m} n^{\nu} \cdot 1 = n^{\nu} m,$$

i.e. the complexity of the decomposed system is not less than that of the initial one L. This should have been expected, since such decomposition only reflects a possibility to present the vector model $\mathbf{Y} = \mathbf{F}(\mathbf{X})$ in a scalar form

$$y_i = f_i(\mathbf{X}), \quad i = 1, \ldots, m, \tag{3.3.6}$$

which, naturally, does not simplify the problem. However, if there are some a priori data on the object that its certain outputs y_i in (3.3.6) are independent from its certain inputs, then due regard of the data in the process of decomposition, as is easy to see, decreases the complexity of the decomposed system, i.e. makes $L_d < L$. Obviously, this is the case when

$$L_d = \sum_{i=1}^{m} n_i^{\nu},$$

where n_i is the number of inputs of the i-th element of the decomposed model ($n_i \leqslant n$), i.e. $L_d < L$.

Let us consider an example of object decomposition into two different elements according to N different methods (Fig. 3.3.3). This means that the set $\{D\}$ consists of N versions, the numbers k and q being different for the different versions of decomposition. Obviously, we can consider a decomposition to be good, if these numbers are minimal. The problem may be put on the basis of (3.3.2) and (3.3.3) as

$$L(k, q) = (n+q)^{\nu} k + k^{\nu}(m+q) \to \min_{\{D\}} \Rightarrow D^*.$$

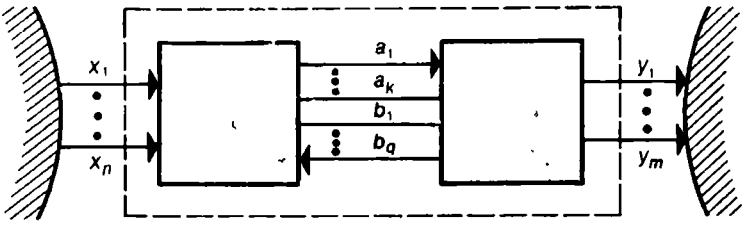

Fig. 3.3.3. An example of decomposition into two elements

The decomposition D^* out of $\{D\}$ is going to be optimal, if it minimizes $L(k, q)$. Let us call D_i an i-th decomposition which is determined by a pair of numbers

$$D_i = \langle k_i, q_i \rangle, \quad (i = 1, \ldots, N).$$

The solution for the problem is $D^* = D_l$, if

$$L(k_l, q_l) = \min_{i=1, \ldots, N} \{L(k_i, q_i)\}.$$

Consequently, the aim of decomposition is first of all to simplify the subsequent synthesis of the object's model by means of 'splitting' it into simpler elements. The process should necessarily involve consideration of a priori data on the structural features of the object.

As a result of decomposition of the object into g elements, the problem to synthesize the model of the object is reduced to g problems to synthesize the models of each element, i.e. to simpler problems.

The process of decomposition may be applied (and sometimes it is expedient) to the resulting elements, thus decomposing them into still smaller components. However, it is clear that there should be a limit to the process, when further 'splitting' of the object starts to aggravate the problem. Fig. 3.3.4*a* shows the number of the elements and their complexity versus the level of decomposition. When the level of decomposition is high, the complexity of each element cannot be decreased very much, which results in an increased complexity of the decomposed object (see Fig. 3.3.4*b*). The optimal decomposition minimizes the complexity of the decomposed system.

Somehow or other, by means of decomposition we manage to reduce the initial $(n \times m)$-port (a 'black box')

to a set of simpler and lesser 'black boxes', so that to be able to determine only their structure at the next stage of the structural synthesis. The process of decomposition proper may be thus reduced to identifying the elements of the object and to finding out the interrelationships between these elements in the object.

3.4. The Structure of the Model

The process of synthesizing the operator **F** of the model of the object $\mathbf{F}^o$ (or the model's elements) may be reduced to determining both the structure St of the model and its parameters $\mathbf{C} = (c_1, \ldots, c_k)$, i.e. the operator of the model **F** according to (I.4.2) may be presented as a pair

$$\mathbf{F} = \langle St, \mathbf{C} \rangle, \tag{3.4.1}$$

and the relationship of inputs and outputs of the model as

$$\mathbf{Y} = \mathbf{F}_{St}(\mathbf{X}, \mathbf{U}, \mathbf{C}), \tag{3.4.2}$$

where $\mathbf{F}_{St}$ is the operator of transformation with the structure St whose parameters are carried out into the variables **C** for convenience. (The index St will be further omitted, since there is no operator without a structure.) The presentation of the transformation operator as (3.4.2) is the result of the parametrization of the model, for which its structure St is to be determined, and the parameters **C** of the model will be unknown so far.

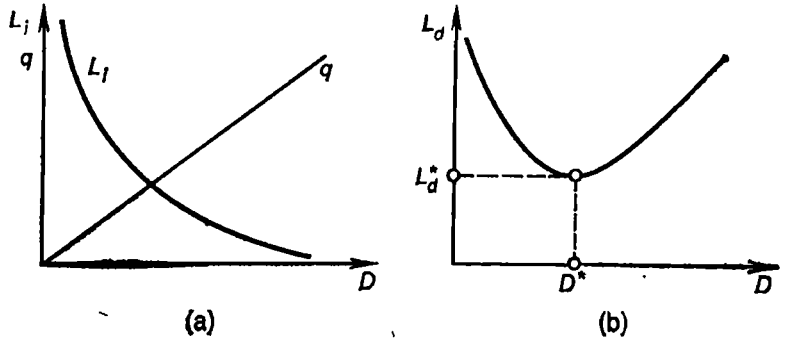

Fig. 3.3.4. Influence of the level of decomposition on the complexity of the elements L_i, their number q (*a*), and the complexity of the decomposed object (*b*)

Let us consider the main categories of the object's model structure: dynamics, nonlinearity, stochasticity, and non-stationarity.

Dynamics. From the point of view of dynamics of the object of control, it is expedient to consider the static and dynamic structures of the object.

We shall call the structure of the object static, if it is enough to apply a model in the form of the function **F** for its description, and dynamic, if any adequate description of the object requires operators with memory (e.g. differential operators, integrators, time lag, etc.).

An example of a static structure is the structure described by decomposition of the output of the object into a certain system of independent linear input functions

$$y = \sum_{i=1}^{k} c_i \varphi_i (\mathbf{X}, \mathbf{U}), \tag{3.4.3}$$

where $\{\varphi_i (\cdot, \cdot)\}$ is a system of functions of inputs, the selection of which defines the selection of a static structure of the object; $c_1, \ldots, c_k$ are the parameters of the object to be determined at the stage of identification (see Chapter 5). Let us recall the example with the cocktail (§ I.5) the model of which was a static structure with the functions $\varphi_j (\mathbf{X}, \mathbf{U}) = u_j$ $(j = 1, 2, 3)$.

An example of a dynamic structure with $n = m = 1$ $(\mathbf{X} = x, \mathbf{Y} = y)$ is a model described by a linear differential equation

$$\frac{d^p y}{dt^p} + a_{p-1} \frac{d^{p-1} y}{dt^{p-1}} + \ldots + a_1 \frac{dy}{dt} + a_0 y = \\ = b_l \frac{d^l x}{dt^l} + \ldots + b_1 \frac{dx}{dt} + b_0 x \tag{3.4.4}$$

or in a more compact form

$$\sum_{i=1}^{p} a_i y^{(i)} = \sum_{j=1}^{l} b_j x^{(j)},$$

where $y^{(i)} = \frac{d^i y}{dt^i}$, $x^{(j)} = \frac{d^j x}{dt^j}$; while a_i $(i = 0, \ldots, p-1)$ and b_j $(j = 0, \ldots, l)$, $a_p = 1$ are the parameters of the continuous model.

In the case of a discrete model, equation (3.4.4) may be written in a recurrent (finite difference) form:

$$y_r = \sum_{i=1}^{p} d_i y_{r-i} + \sum_{j=1}^{l+1} e_j x_{r-j+1}, \qquad (3.4.5)$$

where $d_1, \ldots, d_p$ and $e_1, \ldots, e_{l+1}$ are the parameters of the discrete model, while x_r and y_r are the values of input and output of the model in the r-th time moment. (Time in this case is assumed to be discrete, i.e. $t = r = 1, 2, \ldots$.)

For instance, a model of such an object as a simple pendulum with small amplitudes has the following structure:

$$\frac{d^2y}{dt^2} + a_1 \frac{dy}{dt} + a_0 y = b_0 x(t),$$

where y is the angle between the pendulum and the vertical, and $x(t)$ is the external disturbing force. In case we want a discrete model, the model of the pendulum may be described by a finite difference expression

$$y_r = d_1 y_{r-1} + d_2 y_{r-2} + e_1 x_r.$$

Let us note that the finite difference models are more widely accepted to describe the behaviour of complex systems. We shall apply them broadly further.

Nonlinearity. From the viewpoint of response to external effects, the objects may be conveniently categorized as linear and non-linear. We shall imply by a linear object such an object the reaction of which to the sum of any two external disturbances $\mathbf{X}_1(t)$ and $\mathbf{X}_2(t)$ equals the sum of the reactions to these disturbances, i.e.

$$\mathbf{F}^0(\mathbf{X}_1(t) + \mathbf{X}_2(t)) = \mathbf{F}^0(\mathbf{X}_1(t)) + \mathbf{F}^0(\mathbf{X}_2(t)), \qquad (3.4.6)$$

and it is necessary that the condition $\mathbf{F}^0(0) = 0$ should be observed, i.e. the object should be centred.

The linear static structure is described by a common linear function

$$y = c_1 x_1 + \ldots + c_n x_n,$$

and in its more general form (with linearity by both **X** and **U**) it is

$$y = \sum_{i=1}^{n} c_i x_i + \sum_{j=1}^{q} c_{n+j} u_j. \qquad (3.4.7)$$

Examples of such functions are the properties of cocktail (its taste, strength, and cost) depending linearly on its composition (see § I.5). It is easy to prove that dynamic models (3.4.4) and (3.4.5) are also linear. Suffice it only to substitute them into (3.4.6).

A typical example of a non-linear static structure is our sensation of taste. For instance, when salt is added to unsalted food, its taste first greatly improves, but when more salt is added, it drops sharply (it is indeed 'better too little than too much'). (Let us note that the taste function $f_1(\mathbf{U})$ in the cocktail example discussed in § I.5 was taken to be linear for simplicity's sake which was needed in order to demonstrate how to solve the problem completely. The function is really non-linear.)

When the amplitude is great, the pendulum is another example of a non-linear structure of the following form:

$$\frac{d^2y}{dt^2} + a_1 \frac{dy}{dt} + a_0 \sin y = b_0 x(t).$$

Stochasticity. The stochastic nature of the structure of the object is due to the presence both in the object and in the medium of unobserved but essential factors which may be modelled statistically:

$$\mathbf{Y} = \mathbf{F}(\mathbf{X}, \mathbf{U}, \mathbf{E}(t)), \tag{3.4.8}$$

where $\mathbf{E}(t)$ is a random process to model the existing uncertainty in the object and the medium. This uncertainty may be due to the lack of knowledge of some factors operating in the object and the medium, and to noise in the measurement of the object's inputs and outputs.

Anyway, a stochastic object behaves diversely in the same situations, which is modelled by a random vector $\mathbf{E}(t)$ whose statistical properties should be given. The simplest case is when it is possible to assume

$$\mathbf{Y} = \mathbf{F}(\mathbf{X}, \mathbf{U}) + \mathbf{E}(t). \tag{3.4.9}$$

An example of such a stochastic object is any biological organism behaving, as it is well known, differently in the same conditions. The expression $\mathbf{F}(\mathbf{X}, \mathbf{U})$ in this case describes the object's behaviour strictly determined by the external conditions, and all deviations from the regu-

lar behaviour are included into the 'random disturbance' $\mathbf{E}(t)$.

Non-stationarity. The object's non-stationarity is related to a determined or random change in the object's operator $\mathbf{F}^\circ$ with time. If this change is slow enough, showing as the parameters' drift, it may be disregarded, since the process of correction (adaptation) of the model at every step of control adjusts the model and thus compensates the drift. But if $\mathbf{F}^\circ$ changes rapidly, such non-stationarity should be included in the structure of the model as a time-dependence

$$\mathbf{Y} = \mathbf{F}_t(\mathbf{X}, \mathbf{U}, \mathbf{C}) = \mathbf{F}(\mathbf{X}, \mathbf{U}, \mathbf{C}, t),$$

where time t is a parameter.

For instance, when the parameters of the model change linearly, we can write

$$\mathbf{C} = \mathbf{C}_0 + \mathbf{C}_1 t,$$

where $\mathbf{C}_0$, $\mathbf{C}_1$ are the parameters of the model and its change to be determined at the subsequent stage of identification.

An example of such a non-stationarity is the 'ageing' of the object, i.e. its depreciation with time. This ageing is seen in a certain change of the object's properties, which makes it non-stationary.

Let us note that the listed structural categories do not determine the structure of the model yet, but rather the type of its structure. The process of selecting the structure of the model or its elements may be reduced, as a matter of fact, to selecting a language to describe the object. That description is assumed to be the model.

Here are some examples of this kind of languages: linear algebra for synthesizing linear statistical models of the (3.4.7) type, linear differential equations for synthesizing continuous linear dynamic models of the (3.4.4) type, finite difference equations for synthesizing linear dynamic models of the (3.4.5) type, etc. However, these languages of classical mathematics are suitable for describing only comparatively uncomplicated objects. Complex objects of control require creation of new languages with greater

expressive capabilities. There are many such languages to date.

Let us consider two trends in the field: simulation and semiotic approaches to modelling complex objects.

3.5. Simulation Models

We shall imply by simulation modelling a step-by-step modelling of the object's behaviour with the aid of a computer. This means that certain time moments $t_1, \ldots, t_N$ are fixed, and the state of the model is determined (calculated by the computer) consecutively for each of the moments. This requires a rule (an algorithm) for the model to pass from one state into the next, i.e. a transformation $\mathbf{Y}_i \to \mathbf{Y}_{i+1}$, where $\mathbf{Y}_i$ is the state of the model in the i-th moment.

Let it be, as usual, that the state of the model is presented by the vector $\mathbf{Y} = (y_1, \ldots, y_m)$, i.e. by m numbers, the state of the medium $\mathbf{X} = (x_1, \ldots, x_n)$ by n numbers, the state of the control $\mathbf{U} = (u_1, \ldots, u_q)$ by q numbers. Then the simulation model may be determined by an operator $\mathbf{F}$ affording to find the state of the model in the next moment $(\mathbf{Y}_{i+1})$ knowing the state $\mathbf{Y}_i$ in the preceding moment and the values $\mathbf{X}_{i+1}$ and $\mathbf{U}_{i+1}$:

$$(\mathbf{Y}_i,\ \mathbf{X}_{i+1},\ \mathbf{U}_{i+1}) \xrightarrow{\mathbf{F}} \mathbf{Y}_{i+1}. \qquad (3.5.1)$$

Roughly speaking, the model affords to find out what is going to be tomorrow if we know what there is today. The specificity of simulation modelling is that the state of the object is found out recurrently at every step taking into account the preceding step only. Transformation (3.5.1) may be conveniently written as a recurrent formula

$$\mathbf{Y}_{i+1} = \mathbf{F}(\mathbf{Y}_i,\ \mathbf{X}_{i+1},\ \mathbf{U}_{i+1}), \qquad (3.5.2)$$

where $\mathbf{F}$ is the operator of the simulated changes in the state of the model. This operator presents the simulation model of the object of control with the model's structure St and parameters $\mathbf{C}$ in compliance with (3.4.1).

At the stage of the structural synthesis of the simulation model $\mathbf{F}$ only its structure St and the qualitative contents of its parameters $\mathbf{C} = (c_1, \ldots, c_k)$ are determined,

and the quantitative analysis of the latter is conducted at the stage of identification (see the next chapter).

It is easy to note that simulation models generalize the finite difference model discussed in § 3.4. Let us consider, for instance, a non-linear finite difference model in the form

$$y_r = f(y_{r-1}, \ldots, y_{r-p}, x_r, \ldots, x_{r-s}),$$

where f is a known function. Let us introduce designations

$$\mathbf{Y}_{r-1} = (y_{r-1}, \ldots, y_{r-p}); \ \mathbf{X}_r = (x_r, \ldots, x_{r-s}).$$

Then the equation for y_r may be rewritten in the form

$$y_r = f(\mathbf{Y}_{r-1}, \mathbf{X}_r). \tag{3.5.3}$$

Obviously, this expression is a particular case of (3.5.2). Therefore, the finite difference model is a particular case of the simulation model, just as well as any function is a particular case of an algorithm (since a function is also a rule to calculate its value by the argument, i.e. it is an algorithm).

Simulation models have one more advantage: they make it possible to take into account the unobserved factors **E** of the object, its stochasticity (3.4.8) in the form

$$\mathbf{Y}_{t+1} = \mathbf{F}(\mathbf{Y}_t, \mathbf{X}_{t+1}, \mathbf{U}_{t+1}, \mathbf{E}_{t+1}), \quad t = 1, \ldots, N, \tag{3.5.4}$$

but it is necessary to know how the factors **E** affect the state **Y** of the object. For that we should study the object well and indicate precisely the way the unobserved factors **E** enter the operator of the object so that the data could be included into the operator **F** of the model. To make the model work, it is necessary to know the concrete values of the factors **E** which are, as it is known, unobserved.

This contradiction may be resolved by the so-called Monte Carlo method (or the method of statistical trials). To use it, it is necessary to know some statistical properties of the factor **E** (e.g. its distribution, correlation properties, etc.). Generally, these properties may depend on **Y**, **X**, and **U**. These data being available, we may model the unobserved factor in the form of random sequences

$$\mathbf{E}_1^j, \mathbf{E}_2^j, \ldots, \mathbf{E}_N^j, \quad (j = 1, \ldots, N'), \tag{3.5.5}$$

where the lower index corresponds to a discrete time moment, and the upper to the number of the modelled sequence (the total of N' of such statistically equivalent sequences are modelled). Naturally, neither of the sequences is an accurate realization of reality, but each has the same statistical properties as the real one. This is why sequences (3.5.5) make it possible to study the statistical properties of model (3.5.4). For instance, the behaviour of the model 'on the average' may be described as

$$\bar{\mathbf{Y}}_i = \frac{1}{N'} \sum_{j=1}^{N'} \mathbf{Y}^j_i \quad (i = 1, \ldots, N),$$

where $\mathbf{Y}^j_i$ is a j-th realization of the model behaviour in an i-th time moment:

$$\mathbf{Y}^j_i = \mathbf{F}(\mathbf{Y}^j_{i-1}, \mathbf{X}_i, \mathbf{U}_i, \mathbf{E}^j_i), \quad (i = 1, \ldots, N-1).$$

The dispersion of the output of the model is evidently calculated as

$$D(\mathbf{Y}_i) = \frac{1}{N'-1} \sum_{j=1}^{N'} (\mathbf{Y}^j_i - \bar{\mathbf{Y}}_i)^2.$$

Other statistical properties of the simulated process may be estimated similarly.

This is the essence of the Monte Carlo method: it enables us to estimate the statistical properties of the object's behaviour by means of random 'playing through' the model's behaviour, one realization of the behaviour differing from another by diverse values of the unobserved factor **E**.

3.6. Semiotic Models

The synthesis of any model is a process to formalize the description of the object, i.e. passing over from a natural language to a formalized one (graph, a simulation model, analytical equations, etc.), which is always difficult, especially for complex objects of control. Even now this process is essentially an art. This is due, above all, to the multivalence and 'unclear contour' of the notions of the natural language. However, while describing the

objects of control, this obstacle is easy to overcome by explaining the sense of the notions being used beforehand, i.e. by compiling something like an explanatory dictionary.

The description in the natural language may be formalized in this case, i.e. directly written in a symbolic form by means of a small number of certain symbols—notions immediately related to the notions of the language. Let us call such a model *semiotic* (semiotics is the science of symbolic systems), since it has a clearly pronounced symbolic nature. The advantages of semiotic models are obvious: they may help describe a very large set of objects of control, practically all the objects that can be described in the natural language.

The basis of the semiotic models is the language to describe them. Let us consider one of such languages—the language of binary relations. It is simple enough.

The vocabulary of the binary relations language includes:

—notions (let us denote them by the symbols a_i, $i = 1, 2, \ldots$) that are inherent in the object and introduced on the basis of its description. Roughly speaking, these are the elements of which the described object consists, e.g. a_1 is 'elevator', a_2 is 'button', etc.;

—relations (let us denote them by the symbols r_j, $j = 1, 2, \ldots$) that indicate the relations between the notions. The need in such relations is obvious since it is not enough for an object's description just to describe its elements, it is necessary also to indicate the relations between the elements. These relations are universal; about 200 relations have been identified for the natural language, e.g. r_1—'to be in', r_2—'to be simultaneous with', etc. Let us note that such relations are the specificity of this language which postulates their binary nature meaning that a given relation between two notions either exists or not, and there is no third alternative;

—names, i.e. various designations used in the natural language for the things (let us denote them by the symbols b_k, $k = 1, 2, \ldots$). To introduce the names, there is a special relation r_0—'to have the name'. Apart from such proper names as 'Alex' or 'Mary' these names may, if necessary, specify the notions or make them concrete,

e.g. b_1 is an adjective 'passenger', while b_2 is 'freight', $b_3 = 5$ (the number '5' is a name), etc.

Such is the vocabulary of the binary relations language. It is very convenient and constructive, each of the words has a concrete meaning related to the described object of control. Indeed, notions and names are introduced there as needed, and every object engenders its own individual vocabulary of notions, relations, and names. This is the semantics of the language of binary relations.

Now to the grammar (or, rather, syntax) of that language. Any syntax should indicate precisely and unambiguously how to produce syntactically correct phrases, i.e., it should define the formal signs of syntactically correct phrases (the phrases, though, may be wrong, e.g. 'house is part of room', or senseless, e.g. 'wire simultaneously with elevator', but syntactically correct).

Syntactically correct phrases (A) are:

$A = a_i$, i.e. any notion is a syntactically correct phrase;

$A = (a_i r_j a_k)$, where a_i and a_k are any notions, r_j is any relation apart from r_0 ('to have the name'), i.e. any two notions joined by a relation are also a syntactically correct phrase;

$A = (A_i r_j A_k)$, where A_i and A_k are syntactically correct phrases, and r_j is any relation except for r_0, i.e. two syntactically correct phrases joined by a relation are a syntactically correct phrase;

$A = (A_j r_0 b_k)$, where A_j is a syntactically correct phrase, b_k is any name, i.e. an expression 'A_j has the name b_k' is always a syntactically correct phrase.

There are no other kinds of syntactically correct phrases. This last comment excepts a possibility of emergence of any other phrases. It is very essential for the computer since it does not have human experience to reject incorrect constructions immediately.

Example. Suppose an object of our analysis is a situation described in a phrase of the natural language 'Alex and Mary are in the passenger elevator'. The situation may be presented in the language of binary relations in the following way. The vocabulary:

Notions: $\begin{cases} a_1 - \text{'person'} \\ a_2 - \text{'elevator'} \end{cases}$

Relations: r_0 — 'to have the name'; r_1 — 'to be in'; r_2 — 'to be simultaneously'

Names: b_1 — 'Alex'; b_2 — 'Mary'; b_3 — 'passenger' (adj.).

Let us start with the simplest phrases to be included into the description of the situation:

$A_1 = (a_1 r_0 b_1)$ — 'person named Alex'

$A_2 = (a_1 r_0 b_2)$ — 'person named Mary'

$A_3 = (a_2 r_0 b_3)$ — 'passenger elevator'

$A_4 = (A_1 r_2 A_2)$ — 'person Alex and simultaneously person Mary'

Now the situation to be described may be naturally presented as:

$A_5 = (A_4 r_1 A_3)$ — 'person Alex and simultaneously person Mary are in passenger elevator'.

Additional notion 'person' is not just a bureaucratic style. It is necessary since names may be applied not only to people. For instance, a device 'Mary' is a 'Modified Annihilator with a Radioactive Yield'. Now the same situation 'Alex and Mary are in the passenger elevator' should be written otherwise. The vocabulary should include a new notion a_3—device. Let us form the needed phrases:

$A_6 = (a_3 r_0 b_2)$ — 'device named Mary',

$A_7 = (A_1 r_2 A_6)$ — 'person Alex and simultaneously device Mary'.

And the result is

$A_8 = (A_7 r_1 A_3)$ — 'person Alex and simultaneously device Mary are in passenger elevator'.

Introducing additional brackets, both situations may be described in the symbols of the vocabulary:

$$A_5 = (((a_1 r_0 b_1)\ r_2\ (a_1 r_0 b_2))\ r_1\ (a_2 r_0 b_3)),$$

$$A_8 = (((a_1 r_0 b_1)\ r_2\ (a_3 r_0 b_2))\ r_1\ (a_2 r_0 b_3)).$$

This is achieved by a direct substitution of the expressions for A_1 and A_2 into A_4 which is substituted together with A_3 into A_5, etc. Naturally, nobody is bothered with the adventures of our Alex and Mary in the elevator, that is just a demonstrative example how to describe a situation.

The situation involved is static. The important thing is only the presence of Alex and Mary in the elevator. But if we are interested in the process of the elevator operation, it is necessary to introduce dynamics, i.e. a change of the situation with time. To describe the movement of the elevator, we should introduce another

notion a_4—'floor', names for it: 1, 2, 3 . . ., N, and a relation r_3 'to be at'. Then the upward movement of the elevator is a sequence of situations:

$B_1 = (a_2 r_3 (a_4 r_0 1))$—'elevator is at the first floor',

. . . . . .

$B_N = (a_2 r_3 (a_4 r_0 N))$—'elevator is at the N-th floor'.

In case there are some external effects on the object depending on time, it is necessary to introduce a notion a_5—'time', and the names of time: 1 2, 3, . . . (in some convenient units—minutes, hours, etc.) that change from one situation to another.

Besides, it is necessary to introduce the relation r_4—'engender' which permits to introduce external effects with some given names g_1, . . ., g_L as being 'engendered' at certain time moments: $(a_5 r_0 i)\ r_4\ (a_j r_0 g_i)$, i.e. 'at the i-th time moment a_j acquires name g_i'.

For instance, the program of the elevator movement may be effected by means of the notions d_j—'the j-th button in the elevator' ($j = 1, \ldots, N$) and two names ($L = 2$): g_1—'pressed', g_2—'released'.

The control results in creating certain situations in the object and may be reduced to a change in the names of the notions used to describe the object. The synthesis of control is a compilation of phrases

$$(B_j r_4 (a_l r_0 g_k)), \quad j = 1, \ldots, \qquad (3.6.1)$$

each of them indicating the name g_k acquired by the notion a_l in a situation B_j. For instance, if

$B_j = (a_2 r_3 (a_4 r_0 j))$—'elevator is at the j-th floor', then the following expressions

$(B_N r_4 (d_1 r_0 g_1))$—'at the N-th floor the first button is pressed',

$(B_1 r_4 (d_N r_0 g_1))$—'at the first floor the N-th button is pressed' mean that our 'Alex and Mary are riding the elevator'. It is necessary to add the condition

$(B_j r_4 (d_j r_0 g_2))$—'at the j-th floor, the j-th button is released' which is carried out by the automatics of the elevator.

It is clear that the binary relations language permits to describe symbolically very complex situations in complex systems. However, a symbolic description as it is does not take us too far (apart from, though, reduced notation, which is sometimes of no small importance). A symbolic description of an object should provide a possibility to identify certain features of the object which do not obviously follow from its initial description. Various formal transformations are used for that, which help change the initial description of the object as needed. And that may be done by a computer, since the transformations are formalized.

The transformations are underlied by the algebraic properties of the relations r_i. Thus, the said relation r_2 ('to be simultaneously') has the property of symmetry. It means that $(A_1 r_2 A_2) = (A_2 r_2 A_1)$, i.e. if A_1 is simultaneous with A_2, then A_2 is simultaneous with A_1. The relation r_1 ('to be in') is transitive, i.e. $(A_1 r_1 A_3)$ follows from $(A_1 r_1 A_2)$ and $(A_2 r_1 A_3)$. Indeed, if A_1 is inside A_2, and A_2 inside A_3, then A_1 is inside A_3 (as in the well-known Russian fairy-tale: the needle is in an egg, the egg is in a duck, the duck is in a chest, etc.; therefore, the needle is in the chest, and that is a fairy-tale use of the algebraic property of transitivity).

At present, about 400 of such properties have been identified to help transform an initial description.

Another method to transform such a description of an object is to create new notions. By a new notion we shall mean a situation often repeated in the description of an object and related to the objectives of control. Thus, 'a taxi moving without a passenger' is a situation which worries taxi-drivers very much (this is why they have a slang expression 'idle run').

Generalized are situations that differ only in notions (generalization by notions), relations (generalization by relations), or names (generalization by names). For instance, when describing the operation of the post, it is important to identify regular recipients. They may be individual persons, offices, enterprises, shops, etc. They are all 'correspondence consumers'. It is natural to generalize merging them into a notion 'user'.

When describing diverse properties of an object connected with its relation to measuring instruments, it is natural to combine them and generalize making up the notions of chemical, physical, social properties, etc.

The most widely practised is generalization by names. Is it essential for the elevator control that the names of the passengers are Alex and Mary? Most likely not. Therefore, the description of the elevator operation may be naturally generalized by the names of the passengers. This means to have instead of a cumbersome description 'person Alex and simultaneously person Mary are in elevator' just as informative 'two persons are in elevator'.

All these methods to transform the initial unwieldy

description of the object make it possible to reduce it to a more compact form and, at the next stage, to set about the identification of the unknown parameters in the description (usually they are names) and the synthesis of control (3.6.1), i.e. the determination of such names and time moments for their introduction which permit the model to achieve the objectives of control.

Comments on References to Chapter 3

The problem of using formal approaches for the determination of the structure of the model is rather new. Since a priori information on the object structure is hard to be formalized, the expert methods prevail here almost completely so far. There are no comprehensive text-books yet on the structural synthesis of a wide enough class of the objects of control. There is only one chapter from the book [1] to be recommended.

Comments on the method of expert evaluation was presented in Chapter 2 along with the relevant references.

Simulation models discussed in § 3.5 are the most popular class of models applied to describe the behaviour of complex control systems. Among their advantages is their compatibility with the classical (analytical) models and the Monte Carlo models the latter using the principle of statistical trials. Efficient and sophisticated methods of simulation modelling have already been developed, they make it possible to synthesise models of complex technical, technological [3, 4], and even socio-economic [12, 13] systems. These models help control the complex objects mainly on the basis of the behaviour prognosis for such objects obtained with the aid of simulation.

Initial introduction into simulation models may be provided by [6-8, 15]. The main Monte Carlo methods are basically dealt with in [9] and in more detail in the books [2, 10, 14]. The book [2-5, 12, 13] are recommended for a more thorough study of simulation modelling.

A new class of models appeared recently—the semiotic models discussed in § 3.6. They help formalize objects the behaviour of which may be translated into a language close to natural. Semiotic models in their capabilities are second only to descriptions in the natural language. This allows considering them to be the most promising models to describe complex objects of control. The main features and applications of semiotic models are presented in the brochure [11], while the language of binary relations is tackled with comprehensively in [16], though for information retrieval systems only.

References

1. Rastrigin, L.A., Madzharov, N.E. *Introduction into Identification of Objects of Control*. Energiya, Moscow, 1977 (in Russian).

2. Yermakov, S.M. *The Monte Carlo Method and the Related Problems.* Nauka, Moscow, 1971 (in Russian).
3. Buslenko, N.P. *Mathematical Modelling of Industrial Processes on Numerical Computers.* Nauka, Moscow, 1964 (in Russian).
4. Buslenko, N.P. *Modelling of Complex Systems.* Nauka, Moscow, 1968 (in Russian).
5. Wagner, H.M. *Principles of Operations Research with Applications to Managerial Decisions.* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1969.
6. Moiseev, N.N. *Mathematician Asks Questions.* Znaniye, Moscow, 1974 (in Russian).
7. Moiseev, N.N. Simulation Models. In: *Science and Mankind*: International annual publication. Znaniye, Moscow, 1973 (in Russian, German, etc.).
8. Koryavov, P.P., Sushkov, B.G. *Simulation of Dynamic Processes.* Znaniye, Moscow, 1973 (in Russian).
9. Sobol, I.M. *The Monte Carlo Method.* Mir Publishers, Moscow, 1975.
10. Pollyak, Yu.G. *Probabilistic modelling on computers.* Sov. radio, Moscow, 1971 (in Russian).
11. Pospelov, D.A. *Large Systems: Situational Control.* Znaniye, Moscow, 1975 (in Russian).
12. Naylor, T.H. *Computer Simulation Experiments with Models of Economic Systems.* Wiley, Inc., New York, 1971.
13. Forrester, J.W. *World Dynamics.* Wright-Allen Press, Inc., Cambridge, Mass., 1971.
14. *The Method of Statistical Trials.* Nauka, Moscow, 1962 (in Russian).
15. Shannon, R.E. *Systems Simulation*: The Art and Science. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975.
16. *Information retrieval system* 'BIT'. Naukova Dumka, Kiev, 1968 (in Russian).

Chapter 4 Identification of the Parameters of the Model

Now that the structure St of the operator **F** of the model has been found out, it is necessary to determine the model's unknown parameters $c_1, \ldots, c_k$. As it was mentioned before (in the 'Introduction'), this may be realized depending on the operator **F** in two ways: by identification and by design of experiments with the object. The first approach (identification) helps determine the parameters of the relations between the inputs and outputs of the object (the parameters of the model) in case there is a possibility to observe the changing inputs of the object, i.e. in case of routine maintenance. The second approach (experiments design) is applied when the controlled inputs of the objects do not change with time, and it is necessary to design special experiments to determine the parameters of the model showing the relation between these inputs and the outputs of the object. This chapter deals with the identification of the parameters of the model.

4.1. Formulation of the Problem of Identification

Thus, we implicate by the identification of the model the process of determining its parameters

$$\mathbf{C} = (c_1, \ldots, c_k) \tag{4.1.1}$$

during routine maintenance of the object. The structure of the model is already clear (it was determined at the stage of the structural synthesis):

$$\mathbf{Y} = \mathbf{F}(\mathbf{X}, \mathbf{U}, \mathbf{C}), \tag{4.1.2}$$

i.e. the operator **F** is assumed to be known. This means that there is a known algorithm (rule, instruction) making it possible to determine the state **Y** of the model if the

states X and U of its inputs are given, as well as parameters (4.1.1). These are the parameters to be determined at the stage of identification.

Evidently, it is necessary for the identification to have some data on the change of the inputs and outputs of the object. But the object is not yet controlled (we are just going to produce the control system), therefore the effect of the input U on the output Y cannot be found out at the stage of identification. This simplifies the problem to a certain extent, since instead of model (4.1.2) we shall have the model

$$\mathbf{Y} = \mathbf{F}'(\mathbf{X}, \mathbf{C}) \tag{4.1.3}$$

that does not include the controlled input U (its effect on the output will be considered at the stage of experiment design).

The process of identification makes use of the initial data which fall conveniently into two categories:

—*a priori* data on the structure *St* of the model. This implies that the form of the operator F′ should be either given or determined at the stage of the structural synthesis discussed in Chapter 3. It may be, for instance, an equation, a graph of relations between the elements of the model, etc.;

—*a posteriori* data that are observations of the states of the input X and output Y of the object in the course of its routine maintenance, i.e. the information

$$I = \langle \mathbf{X}_i, \mathbf{Y}_i \rangle, \quad i = 1, \ldots, N, \tag{4.1.4}$$

where i is the number of the time moment t_i when the values of X and Y were registered, i.e. $\mathbf{X}_i = \mathbf{X}(t_i)$, $\mathbf{Y}_i = \mathbf{Y}(t_i)$, where $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ are the functions describing the behaviour of the input and output of the object in the course of its normal operation in the medium. Usually, the time moments t_i are spread uniformly over the observation period $[0, T]$, i.e. $t_i = \tau(i - 1)$, where τ is the interval between observations (4.1.4), i.e. $\tau = T/(N - 1)$.

Consequently, the initial data necessary for the identification come into a pair

$$\langle St, I \rangle, \tag{4.1.5}$$

i.e. model structure (4.1.3) and observations (4.1.4). The process of identifying the parameters of the model thus may be reduced to the determination of parameters (4.1.1) by the initial data (4.1.5), i.e.

$$\mathbf{C} = \varphi\,(St,\ I), \tag{4.1.6}$$

where φ is the identification algorithm showing how to find the parameters $\mathbf{C}$ if St and I are known.

This chapter is going to deal with various algorithms φ. They may be subdivided into two large classes: adaptive and non-adaptive.

An *adaptive algorithm of identification* implies an algorithm allowing to make the values of identified parameters of the model the more accurate the more there is additional information on the operation of the object. Suppose there are certain values of identified parameters at an i-th step of adaptive identification. Let us denote them by the index i:

$$\mathbf{C}_i = (c_1^i,\ \ldots,\ c_k^i).$$

Suppose we obtain some additional data, i.e. a pair of observations of the input and output of the object in the $(i+1)$-th time moment:

$$I_{i+1} = \langle \mathbf{X}_{i+1},\ \mathbf{Y}_{i+1} \rangle. \tag{4.1.7}$$

Obviously, these data should somehow change (correct) the existing values of $\mathbf{C}$ and make it possible to get $\mathbf{C}_{i+1}$, i.e. more accurate values of the parameters. The relationship between $\mathbf{C}_i$ and $\mathbf{C}_{i+1}$ is defined by the adaptive identification algorithm:

$$(\mathbf{C}_i,\ I_{i+1}) \xrightarrow{\widetilde{\varphi}_a} \mathbf{C}_{i+1}$$

or in the common recurrent form

$$\mathbf{C}_{i+1} = \widetilde{\varphi}_a\,(\mathbf{C}_i,\ I_{i+1}). \tag{4.1.8}$$

Here $\widetilde{\varphi}_a$ is the algorithm for the adaptive identification which determines the subsequent values of the parameters on the basis of the new data (I_{i+1}) and the preceding values of the parameters $(\mathbf{C}_i)$. Therefore, the adaptation represents a method for obtaining 'new knowledge' by way of correcting 'old knowledge' on the basis of new data.

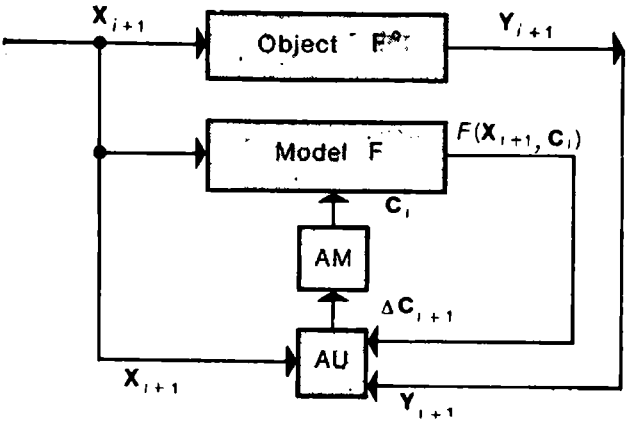

Fig. 4.1.1. Diagram of adaptive identification

Algorithm (4.1.8) may be put in a more convenient form

$$C_{i+1} = C_i + \varphi_a (F (X_{i+1}, C_i), I_{i+1}), \quad (4.1.9)$$

where φ_a is the operator of the adaptive identification.

In case the adaptive identification method is realized in real time, then it is called *the method of the self-adjusting model.* The method is presented diagrammatically in Fig. 4.1.1. The input X of the object is fed to the input of the model. The data on the state of the object Y, model F (X, C), and medium X are fed into the adjustment unit (AU) that produces a correction signal $\Delta C = \varphi_a (F (X, C), I)$ changing the parameters of the model by means of the actuating mechanism (AM).

Apparently, the adaptive identification algorithms are not at all necessarily realized in real time. The role of the object, if that is the case, is performed by the information I (4.1.4) that is supplied to the algorithm for adaptive identification from the memory by portions $\langle X_i, Y_i \rangle$ $(i = 1, \ldots, N)$.

Unlike the adaptive, the non-adaptive identification algorithm permits obtaining the sought parameters C all at once making use of the whole of information I (4.1.4) as opposed to their gradual approximation (we shall consider further various versions of such an algorithm). In case the information I is known, the problem of identification may be solved both by adaptive and non-

adaptive methods. It may seem at the first glance that the non-adaptive algorithm is always better than the adaptive, but one should be cautious. Each of them has its advantages and disadvantages.

The non-adaptive algorithm produces all the identified parameters C at once, but it is more sophisticated and requires for its realization, as a rule, powerful general purpose computers.

The adaptive algorithm is less sophisticated: it is easier to program and debug. Besides, it may be efficiently realized in specialized computers. It is usually applied for identification in the objects with changing characteristics (for instance, drifting parameters) so that it operates in the mode of a self-adjusting model. However, it may also be employed for the object identification over the entire volume of information I. When this is the case, efficient identification of parameters C requires, as a rule, many 'runs' of the information I through the adaptive algorithm.

4.2. Identification of Static Objects

Let us consider model (4.1.3) of the static object, where the form of the function $\mathbf{F}'$ was already determined at the stage of the structural synthesis (see Chapter 3). The function may be analytical or algorithmical; in the latter case it is a set of instructions to calculate the function for all possible values of the arguments X and C.

Let us consider the output of the object to be unidimensional ($m = 1$), i.e. $\mathbf{Y} = y$. This makes it possible to write model (4.1.3) in the form

$$y = f(\mathbf{X}, \mathbf{C}), \tag{4.2.1}$$

where f is a scalar function specified at the stage of the structural synthesis.

The validity of such an approach follows from the obvious decomposition of model (4.1.3), where

$$\mathbf{Y} = (y_1, \ldots, y_m), \quad \mathbf{F}' = (f_1, \ldots, f_m),$$

into m models of the (4.2.1) form, i.e. from decomposing the model like it is shown in Fig. 3.3.2. This implies that

the ability to identify the parameters of model (4.2.1) with only one output allows also identification of the models with many outputs by means of m-fold reiteration of the identification of model (4.2.1).

Initial data (4.1.4) in this case are

$$I = \langle \mathbf{X}_i, y_i \rangle, \quad i = 1, \ldots, N, \tag{4.2.2}$$

where y_i is the response of the object in the i-th time moment to the input $\mathbf{X}_i = (x_1^i, \ldots, x_n^i)$, $y_i = y(t_i)$, $x_j^i = x_j(t_i)$.

Let us consider first the non-adaptive algorithm for identification. In order to determine the parameters $\mathbf{C}$ by the non-adaptive method, let us substitute data (4.2.2) into (4.2.1). The result is a system of N equations in k unknowns:

$$f(\mathbf{X}_1, c_1, \ldots, c_k) = y_1,$$

$$\ldots\ldots\ldots\ldots$$

$$f(\mathbf{X}_N, c_1, \ldots, c_k) = y_N,$$

which is more convenient to have in the form

$$f(\mathbf{X}_i, \mathbf{C}) = y_i, \quad i = 1, \ldots, N. \tag{4.2.3}$$

Obviously, the requirement $N \geqslant k$ has to be met. Otherwise, the number of the unknowns will prove to be less than the number of the equations and system (4.2.3) will not have an unambiguous solution.

Consequently, the problem of non-adaptive identification of the static object may be reduced to the problem of solving the system of equations (4.2.3). The system may have two essential properties which will bring about difficulties in solving it: incompatibility and transcendence. It may be incompatible when $N > k$ (the number of equations is more than that of unknowns), and it may be transcendent due to the nature of the function f.

Let us start with incompatibility ($N > k$). We can solve such an incompatible system of equations (4.2.3) by the least-squares method, i.e. by minimizing the total discrepancy between the right and left parts of the equations. For the purpose, we shall make a function of the total discrepancy as the sum of the squares of the dis-

crepancies for each of the functions:

$$Q(\mathbf{C}) = \sum_{i=1}^{N} [f(\mathbf{X}_i, \mathbf{C}) - y_i]^2. \qquad (4.2.4)$$

Obviously, this function is non-negative, and it is equal to zero if the right- and left-hand sides in each of the equations are the same. And the closer the right-hand sides in the equations of system (4.2.3) to their left-hand sides, the less the value of discrepancy function (4.2.4). This is the reason why we can consider the solution of system (4.2.3) to be the values of the parameters

$$\mathbf{C}^* = (c_1^*, \ldots, c_k^*) \qquad (4.2.5)$$

at which the discrepancy function is minimal, i.e.

$$Q(\mathbf{C}^*) = \min_{\mathbf{C}} Q(\mathbf{C}).$$

Therefore, to solve an incompatible system of equations (4.2.3), it is sufficient to minimize the function of total discrepancy (4.2.4). This may be put as

$$Q(\mathbf{C}) \to \min_{\mathbf{C}}, \qquad (4.2.6)$$

which is to be read 'minimize the function Q (C) with respect to C'.

Suppose the solution to the problem is C*, for which the evident inequality

$$Q(\mathbf{C}^*) \leqslant Q(\mathbf{C}) \qquad (4.2.7)$$

holds true for any C (let us recall that C is vector (4.1.1), i.e. a sequence of k components $c_1, \ldots, c_k$).

Now the problem of identifying static systems has been reduced to a problem of minimizing a certain specially constructed function of many variables. How is this problem to be solved? Computing mathematics developed for that a number of efficient methods which take account of specific features of the minimized function Q (C).

Further on we shall often have to solve minimization problems (not necessarily concerning discrepancy functions), therefore, the most efficient existing now techniques for their solution are given in the 'Supplement'. The problem of identifying static objects may be solved by means of applying one of them.

Let us note that in identifying complex systems overall discrepancy function (4.2.4) may prove to be very complicated, and there is no chance in trying to minimize it by the 'old-time' method, i.e. without computers. This is bound to fail. It is necessary to apply here modern computers and the efficient search for methods described in the 'Supplement'.

Let us consider the simplest possible case when the function f has the form

$$f(\mathbf{X}, \mathbf{C}) = \sum_{j=1}^{k} c_j \varphi_j(\mathbf{X}), \tag{4.2.8}$$

where $\varphi_j(\mathbf{X})$ $(j = 1, \ldots, k)$ is a specified system of functions of many variables. The specific feature of such a structure is linearity of the function with regard to the sought parameters $c_1, \ldots, c_k$. This linearity reduces the minimization problem (4.2.6) to the solution of a system of linear algebraic equations. It may be solved easily enough by the common methods.

The simple form of function (4.2.8) allows solving the problem to minimize (4.2.6) by making the partial derivatives of the function $Q(\mathbf{C})$ equal to zero, i.e.

$$\frac{\partial Q(\mathbf{C})}{\partial c_l} = 0, \quad l = 1, \ldots, k. \tag{4.2.9}$$

Since f in (4.2.8) is a linear function with regard to C, then $Q(\mathbf{C})$ in (4.2.4) is a quadratic function, and that predetermines linearity of the system of equations (4.2.9). In fact, substituting (4.2.8) into (4.2.4) and differentiating, we get instead of (4.2.9) after some obvious transformations

$$\sum_{j=1}^{k} c_j \psi_{jl} - \eta_l = 0, \quad l = 1, \ldots, k, \tag{4.2.10}$$

where

$$\begin{aligned} \psi_{jl} &= \sum_{i=1}^{N} \varphi_j(\mathbf{X}_i)\, \varphi_l(\mathbf{X}_i), \\ \eta_l &= \sum_{i=1}^{N} y_i \psi_l(\mathbf{X}_i), \quad j, l = 1, \ldots, k. \end{aligned} \tag{4.2.11}$$

Evidently, (4.2.10) is a system of linear algebraic equations with regard to parameters (4.1.1) to be identified.

The system may be solved easily by the common methods (e.g. the Gaussian reduction, Cramer's rule, etc.).

But there should be no mistaking here. The price of easiness is a serious simplification: it has been assumed that the behaviour of a complex object may be presented in the form (4.2.8), i.e. as if there exist the functions $\varphi_i(\mathbf{X})$ describing accurately enough how the object behaves. This assumption seldom holds for a complex object. For instance simulation models and situational ones cannot be identified by means of this method, because the attempts to fit such functions usually fail. However, this approach may serve in a number of cases as the first approximate method to obtain a model of the complex object.

But let us turn back to the system of equations (4.2.10) that is described in detail below. Let us consider the matrix of the system ($k \times k$)

$$\Phi = \left\| \begin{matrix} \psi_{11} & \cdots & \psi_{1k} \\ \cdots & \cdots & \cdots \\ \psi_{k1} & \cdots & \psi_{kk} \end{matrix} \right\|. \qquad (4.2.12)$$

This matrix is known as the Fischer information matrix. It is symmetrical, since $\psi_{ij} = \psi_{ji}$. For an unambiguous solution, it is required that the determinant $|\Phi|$ of the matrix should not be equal to zero, i.e.

$$|\Phi| \neq 0. \qquad (4.2.13)$$

The determinant is equal to zero in two cases: when $k > N$, i.e. when there are not enough measurements, and when the variability of the input $\mathbf{X}$ is insufficient, i.e. when more than $(N - k)$ states $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are linearly dependent. This may be dealt with in two ways: either by increasing the number N of observations, or (in case this does not help because of linear dependence) by decreasing the dimension of the observed input $\mathbf{X}$ of the object, i.e. by lowering the number k.

Now let us consider adaptive identification for the static object. Let $\mathbf{C}_i$ be the values of the identified parameters at the i-th step of the adaptive identification. Suppose we get new data $I_{i+1} = \langle \mathbf{X}_{i+1}, y_{i+1} \rangle$. These data should help substitute $\mathbf{C}_i$ by C_{i+1}, i.e. be the

source for correcting the parameters:

$$\mathbf{C}_{i+1} = \mathbf{C}_i + \Delta\mathbf{C}_{i+1}.$$

The problem therefore is to determine $\Delta\mathbf{C}_{i+1}$ with the aid of I_{i+1}. Let us write for the purpose the local discrepancy between the outputs of the model and the object at the moment $(i + 1)$:

$$q_{i+1}(\mathbf{C}_i) = f(\mathbf{X}_{i+1}, \mathbf{C}_i) - y_{i+1}. \qquad (4.2.14)$$

Apparently, the value of $\Delta\mathbf{C}_{i+1}$ should be such that to diminish the square of this discrepancy. That may be easily achieved by making the step $\Delta\mathbf{C}_{i+1}$ 'antigradient', i.e.

$$\Delta\mathbf{C}_{i+1} = -\alpha_{i+1}\nabla q^2_{i+1}(\mathbf{C}_i). \qquad (4.2.15)$$

Here α_{i+1} is a certain positive coefficient (we shall discuss it below), and ∇ is 'nabla', the symbol of the operator of the gradient of the function following the symbol (see the 'Supplement'):

$$\nabla = \left(\frac{\partial}{\partial c_1}, \ldots, \frac{\partial}{\partial c_n}\right).$$

(Let us recall that the gradient is the vector oriented in the direction of the maximal increase of the function. This is why the step is taken in the opposite antigradient direction of the function $q^2(\mathbf{C})$ since that is the function to be minimized.)

Let us find the gradient of the function $q^2(\mathbf{C})$:

$$\nabla q^2(\mathbf{C}) = \left(\frac{\partial q^2(\mathbf{C})}{\partial c_1}, \ldots, \frac{\partial q^2(\mathbf{C})}{\partial c_k}\right) =$$

$$= 2q(\mathbf{C})\left(\frac{\partial q(\mathbf{C})}{\partial c_1}, \ldots, \frac{\partial q(\mathbf{C})}{\partial c_k}\right) = 2q(\mathbf{C})\nabla q(\mathbf{C}).$$

Substituting here (4.2.14), we get

$$\nabla q^2(\mathbf{C}) = 2q(\mathbf{C})\nabla_{\mathbf{C}} f(\mathbf{X}, \mathbf{C}).$$

Here the subscript $\mathbf{C}$ of ∇ denotes the gradient with respect to the parameters $\mathbf{C}$. This should be taken note of since the function also depends on $\mathbf{X}$.

Consequently, we get the following expression for correcting the parameters at the $(i + 1)$-th step:

$$\Delta\mathbf{C}_{i+1} = -2\alpha_{i+1}q_{i+1}(\mathbf{C}_i)\nabla_{\mathbf{C}} f(\mathbf{X}_{i+1}, \mathbf{C}_i),$$

where

$$\nabla_C f(\mathbf{X}_{i+1}, \mathbf{C}_i) = \left(\frac{\partial f(\mathbf{X}_{i+1}, \mathbf{C})}{\partial c_1}, \ldots, \frac{\partial f(\mathbf{X}_{i+1}, \mathbf{C})}{\partial c_k} \right) \Big|_{\mathbf{C}=\mathbf{C}_i}. \tag{4.2.16}$$

This expression is easily calculated if the analytical expression for the function f is known. For instance, substituting f from (4.2.8), we get

$$\nabla_C f(\mathbf{X}_{i+1}, \mathbf{C}_i) = (\varphi_1(\mathbf{X}_{i+1}), \ldots, \varphi_k(\mathbf{X}_{i+1})).$$

However, the function f for complex objects is specified usually only as an algorithm, so the calculation of the partial derivatives (4.2.16) is not that easy. It is necessary then to make use of the finite difference estimates of the derivatives (see the 'Supplement')

$$\widehat{\frac{\partial f(\mathbf{X}, \mathbf{C})}{\partial c_l}} = \frac{1}{2g}[f(\mathbf{X}, \mathbf{C} + g\mathbf{e}_l) - f(\mathbf{X}, \mathbf{C} - g\mathbf{e}_l)], \tag{4.2.17}$$

$$\widehat{\frac{\partial f(\mathbf{X}, \mathbf{C})}{\partial c_l}} = \frac{1}{g}[f(\mathbf{X}, \mathbf{C} + g\mathbf{e}_l) - f(\mathbf{X}, \mathbf{C})], \tag{4.2.18}$$

where $g > 0$ is the basis of the estimate, and $\mathbf{e}_l$ is the unit vector directed along the l-th axis. Apparently, when $g \to 0$, these estimates tend to the accurate value of the partial derivative.

Now to selection of the parameter α. The correction is going to be small at its low values. But it cannot be too large since the discrepancy will be large too. The trend of the value of discrepancy q^2 versus α is shown in Fig. 4.2.1. Clearly, when $\alpha = \alpha^*$, the discrepancy is minimal and equal to zero. We can make use of that to determine the optimal value of α. It is sufficient for that to solve the problem of monoparametrical minimization

$$q^2(\mathbf{C} - 2\alpha q(\mathbf{C}) \nabla_C f(\mathbf{X}, \mathbf{C})) \to \min_{\alpha > 0}, \tag{4.2.19}$$

i.e. to minimize the square of the discrepancy with regard to the parameter α. However, knowing that the minimal value of the discrepancy, when there is no noise, is

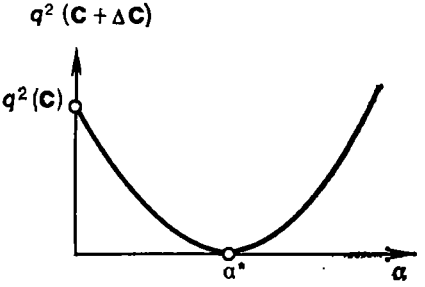

Fig. 4.2.1. To determination of the parameter α

equal to zero, the problem may be solved just making the discrepancy equal to zero:

$$q\,(\mathbf{C} - 2\alpha q\,(\mathbf{C})\nabla_{\mathbf{C}} f\,(\mathbf{X}, \mathbf{C})) = 0. \qquad (4.2.20)$$

The solution to this equation will give the optimal value α^*. For the case of linear function (4.2.8), we get from (4.2.20) after obvious transformations

$$\alpha^* = {}^1/_2 \left[1 + \sum_{j=1}^{n} (x_j^{i+1})^2\right]^{-1},$$

where x_j is the j-th coordinate of the vector of input $\mathbf{X}$. In more involved cases, when the model f is determined as an algorithm (e.g. in the case of simulation models), the optimal value α^* of the parameter should be found while solving the problem of monoparametrical minimization (4.2.19) for which various methods are given in the 'Supplement'.

Now let us come over to the identification of dynamic models.

4.3. Identification of Dynamic Models

The dynamic object differs from the static one in that its state $\mathbf{Y}_i$ is determined by not only the state of its inputs $\mathbf{X}_i$, but also by its own state $\mathbf{Y}_{i-1}$ in the preceding time moment. This is namely the 'memory' of the object which makes it inertial (in the narrow mechanical sense) and dynamic (in the wider sense). Therefore, the discrete model of the dynamic object at the stage of identification may be put in its general form as

$$\mathbf{Y}_i = \mathbf{F}\,(\mathbf{Y}_{i-1}, \mathbf{X}_i, \mathbf{C}), \qquad (4.3.1)$$

where **F** is the function determined at the stage of the structural synthesis, which may be specified analytically and/or algorithmically (the latter form is more common for complex objects of control). The problem of identification is to determine the parameters $\mathbf{C} = (c_1, \ldots, c_h)$ of model (4.3.1) according to observations I of the input and output of the object (see (4.1.4)).

Let us note that the difference in expressions (4.3.1) and (4.1.2) is that **F** is an operator in (4.1.2) while it is a function in (4.3.1) (the absence of control **U** in (4.3.1) is inessential since it is an input just as good as **X** and may be joined with **X** at the stage of identification—naturally, if it ever changes in the mode of normal maintenance of the object).

Let us consider a model with only one output ($m = 1$), as in the preceding section 4.2, which does not, as is known, make the implications less general. To make it simple, suppose the object has only one input ($n = 1$) (further on, it will be shown how this limitation may be lifted). Initial data (4.2.2) on the operation of the object takes the form

$$I = \langle x_i, y_i \rangle, \quad i = 1, \ldots, N, \tag{4.3.2}$$

where x_i and y_i are the values of the input and the output of the object in the i-th time moment t_i, i.e. $x_i = x(t_i)$, $y_i = y(t_i)$.

If the state **Y** of the dynamic object is determined by its output at the present and in p preceding time moments $t_i, \ldots, t_{i-p}$, then such an object is called the dynamic object of the p-th order. Obviously, the state of that object is determined by the values of its output, i.e. by the vector

$$\mathbf{Y}_i' = (y_i, y_{i-1}, \ldots, y_{i-p}). \tag{4.3.3}$$

Now model (4.3.1) of the dynamic object of the p-th order may be presented as

$$y_i = f(\mathbf{Y}_{i-1}', \mathbf{X}_i', \mathbf{C}), \tag{4.3.4}$$

where f is the scalar function specified at the stage of the structural synthesis, vector $\mathbf{Y}_i'$ has been determined in (4.3.3), and $\mathbf{X}_i'$ is the vector of the object disturbance:

$$\mathbf{X}_i' = (x_i, x_{i-1}, \ldots, x_{i-l}) \tag{4.3.5}$$

relating the behaviour of the object with the preceding l states of the medium. The numbers p and l characterize the structure of the object and should be determined at the stage of the structural synthesis of the model.

Let us first consider the non-adaptive method of identification. In this case it is evident that the parameters $\mathbf{C}$ should be such that the total discrepancy between the outputs of the model and the object

$$Q(\mathbf{C}) = \sum_{i=s}^{N} [f(\mathbf{Y}'_{i-1}, \mathbf{X}'_i, \mathbf{C}) - y_i]^2 \qquad (4.3.6)$$

would be minimal, where y_i, $\mathbf{Y}'_{i-1}$, $\mathbf{X}'_i$ are determined by observations (4.3.2), while

$$s = \max\,(p + 2, l + 1), \qquad (4.3.7)$$

i.e. it is the maximal of the two numbers $(p + 2)$ and $(l + 1)$. Discrepancy (4.3.6) is summed up starting with s, and not with l, because it is necessary to have the s preceding values of x_i and y_i.

Now the problem of identifying the parameters $\mathbf{C}$ is reduced to minimization of function (4.3.6):

$$Q(\mathbf{C}) \to \min_{\mathbf{C}}.$$

The problem may be solved by one of the search methods of optimization discussed in the 'Supplement'. The obtained values of the parameters providing the minimum for the function of total discrepancy (4.3.6) are the solution to the problem of identification for the dynamic model described by (4.3.4).

For instance, the total function of discrepancy (4.3.6) for the linear dynamic object the model of which is described by (3.4.5) has the form

$$Q(\mathbf{C}) = \sum_{i=s}^{N} \left[\sum_{r=1}^{p} c_r y_{i-r} + \sum_{j=1}^{l+1} c_{p+j} x_{i-j+1} - y_i \right]^2,$$

where the number of the parameters to be identified is $k = p + l + 1$. Apparently, this is a quadratic form whose minimization is reduced to solving the system of linear algebraic equations:

$$\frac{\partial Q(\mathbf{C})}{\partial c_z} = 0, \quad z = 1, \ldots, k.$$

Now let us generalize this approach over the objects with many inputs ($n \geqslant 1$): $\mathbf{X} = (x_1, \ldots, x_n)$. This is the case when $\mathbf{X}_i'$ in expression (4.3.4) should be understood as the values of the input $\mathbf{X}$ in all l preceding time moments:

$$\mathbf{X}_i' = (\mathbf{X}_i, \mathbf{X}_{i-1}, \ldots, \mathbf{X}_{i-l}),$$

where $\mathbf{X}_r$ $(r = i, i-1, \ldots, i-l)$ is the value of the input $\mathbf{X}$ of the object in the r-th time moment t_r. The initial information in this case has the form (4.2.2).

Consequently, the multiple nature of the input of the object does not complicate the algorithm to solve the problem of identification in principle, but rather makes it more cumbersome. Thus, if there exists a linear dependence (4.3.4) with respect to $\mathbf{C}$, the number of the parameters to be identified goes up from $k = p + l + 1$ (for the unidimensional input) to $k = p + n(l + 1)$ (for the n-dimensional input $\mathbf{X}$).

Now let us consider the adaptive method of identification. It involves minimization of the square of the local discrepancy of the outputs of the model and the object at each step:

$$q^2(\mathbf{C}_{i-1}) = [f(\mathbf{Y}_{i-1}', \mathbf{X}_i', \mathbf{C}_{i-1}) - y_i]^2. \qquad (4.3.8)$$

It may be carried out by means of selecting an appropriate correction $\Delta\mathbf{C}_i$ of the parameters $\mathbf{C}$ as it was shown in the preceding section 4.2, but applying it to (4.3.8):

$$\Delta\mathbf{C}_i = -\alpha_i \nabla_{\mathbf{C}} q^2(\mathbf{C}_{i-1}),$$

where ∇ is the same gradient operator, and the parameter α_i is to be selected so that to minimize the local discrepancy (4.3.8):

$$q^2(\mathbf{C}_{i-1} - \alpha_i \nabla_{\mathbf{C}} q^2(\mathbf{C}_{i-1})) \to \min_{\alpha_i > 0} \Rightarrow \alpha_i^*.$$

In linear case (3.4.5) and when there is no noise, we get

$$\nabla_{\mathbf{C}} q^2(\mathbf{C}_{i-1}) = 2q(\mathbf{C}_{i-1})(y_{i-1}, \ldots, y_{i-p-1}, x_i, \ldots, x_{i-l});$$

$$\alpha^*{}_i = {}^1/_2 \left(\sum_{j=1}^{l} x^2{}_{i-j} + \sum_{s=1}^{p+1} y^2{}_{i-s} \right)^{-1}$$

is the optimal value of the parameter α at which the discrepancy $q(\mathbf{C}_{i-1} + \Delta\mathbf{C}_i)$ is equal to zero.

Let us note in conclusion to this chapter that the process of identification of the parameters $c_1, \ldots, c_k$ of the model of the controlled object always involves the procedure of minimization of 'unbalance' (discrepancy) between the model and the object which is calculated as the difference in the reaction (of the outputs) of the model and the object to the same input. The minimization may be carried out by various search methods (some of which are described in the 'Supplement'). The procedure ends in the determination of the parameters of the model $\mathbf{C}^* = (c_1^*, \ldots, c_k^*)$ most suiting the available data (4.1.4) on the operation of the object. That does not at all mean, however, that the model is adequate to the object. That is only the best approximation for the selected structure of the model. The problem of adequacy should be solved at the stage of selecting the structure of the model and judged by the value of the minimal discrepancy (4.2.4) or (4.3.6). The more the amount of a priori information used when the model structure is determined, the more chances to produce an adequate model.

Comments on References to Chapter 4

Identification as the process to determine the parameters of the model according to observations in the mode of normal operation of the object has been investigated quite intensely. To get introduced into the problem of identification, the reader is recommended the books [1, 2]. Identification of technological objects is the most important practical problem of the control techniques today, and therefore it is given much attention to. Methods to identify such objects are described in [3-5]. The monograph [6] and book [10] present the main identification methods and some of their applications in full enough (though not simple) detail.

The adaptive identification considered in § 4.2 is almost always the most efficient approach to solving practical problems of identification. The publications [7, 8] deal with it specifically, and the book [9] presents a very interesting application of the adaptive identification to solving the problems of control of accuracy in rolled pipe production. It was realized in industry, and the authors were awarded the State Prize in 1976 for it.

References

1. Raibman, N.S. *What is Identification*? Nauka, Moscow, 1981 (in Russian).
2. Rastrigin, L.A., Madzharov, N.Ye. *Introduction into Identification of Objects of Control.* Energiya, Moscow, 1977 (in Russian).

3. Balakirev, S.V., Dudnikov, E.G., Tsirlin, A.M. *Experimental Determination of Dynamic Characteristics of Industrial Objects.* Energiya, Moscow, 1967 (in Russian).
4. Ordintsev, V.M. *Mathematical Description of the Objects of Automation.* Mashinostroenie, Moscow, 1966 (in Russian).
5. Raibman, N.S., Chadeev, V.M. *Design of the Models of Production.* Energiya, Moscow, 1975 (in Russian).
6. Eykhoff, P. *System Identification Parameter and State Estimation.* Wiley, Inc., New York, 1974.
7. Raibman, N.S., Chadeev, V.M. *Adaptive Models in the Systems of Control.* Sov. radio, Moscow, 1966 (in Russian).
8. Raibman, N.S., Chadeev, V.M. *Adaptive Systems to Control Technological Processes: Methods.* Institute of the Problems of Control, Moscow, 1972 (in Russian).
9. Danilov, F.A., Raibman, N.S (Eds.). *Adaptive Control of Accuracy in Rolled Pipe Production.* Metallurgiya, Moscow, 1973 (in Russian).
10. Graupe, D. *Identification of Systems.* Robert E. Krieger Publishing Co., Huntington, N.Y., 1976.

Chapter 5 Experiment Design

The stage of designing experiments with the object emerges first of all due to the necessity to determine the parameters of the influence on the output **Y** of those inputs of the object which do not vary in its normal operation, i.e. without control. Obviously, they are primarily the controlled inputs **U** determined at the stage of the structural synthesis. That was the stage when the dependence

$$\mathbf{Y} = \mathbf{F}(\mathbf{X}, \mathbf{U}, \mathbf{C}) \tag{5.0.1}$$

was obtained, where **F** is the specified operator of the model, and $\mathbf{C} = (c_1, \ldots, c_k)$ contains the unknown parameters of the object to be determined at the stage of the experiment design.

Another aim of the experiment design is related to the determination of the structure of the controlled object's model. This is the case when it is necessary to design such an experiment that the obtained data would make it possible to select the best model among a given set of competing ones and to determine its parameters at that. The need in such an undesirable thing as experiments with the object (let us recall that we have to produce a system of control for the complex object which 'does not like' experiments) is due to the fact that sometimes there is no other way out.

Mathematically, the inevitability of experiments with the object results from the following simple argumentation. To determine the parameters $\mathbf{C} = (c_1, \ldots, c_k)$ of model (5.0.1), it is necessary to have the initial data on the state of the inputs and outputs of the object as sets of threes

$$I = \langle \mathbf{X}_i, \mathbf{U}_i, \mathbf{Y}_i \rangle, \; i = 1, \ldots, N. \tag{5.0.2}$$

But the input **U** at the stage of synthesizing the control system does not change, i.e. $\mathbf{U}_i = \text{const}$, because the

object is not controlled yet. That excludes a possibility to determine the parameters characterizing the connection between the input U and the output Y at the stage of identification. To determine the parameters, it is necessary to vary the input U and register the resulting behaviour of the object, i.e. to experiment with it.

For simplicity of notation, let us consider the case

$$\mathbf{Y} = \mathbf{F}(\mathbf{U}, \mathbf{C}). \qquad (5.0.3)$$

The general case (5.0.1) is easily obtained from (5.0.3) by means of additional registration (recording) the state X of the medium in the process of the experiments.

The problem of experiment design lies in such a minimal change of the input U at which the registered data Y would make it possible to estimate the values of the parameters C best. Let us consider this problem in more detail.

5.1. The Basic Concepts of the Theory of Experiment Design

The theory of experiment design emerged as a part of mathematical statistics and therefore it developed the terms distinct from the terminology of the theory of control (that should be borne in mind while reading literature).

We shall imply by the object of design a multiport with q controlled inputs $\mathbf{U} = (u_1, \ldots, u_q)$. The form (structure) of the operator F (5.0.3) has been already determined—at the stage of structural synthesis of the model considered in Chapter 3. It may have a stochastic structure of the (3.4.9) kind.

To solve the problem of experiment design, we should organize such a change in the inputs U of the object that it would be possible to determine efficiently the parameters C of the model and, if necessary, to correct the structure St of the model F.

A diagram of the process of experiment design is shown in Fig. 5.1.1. It reminds the arrangement of the control system discussed in the 'Introduction' (see Fig. I.2.1). This is not surprising since the experiment design is really a control of the object, but the objectives of that

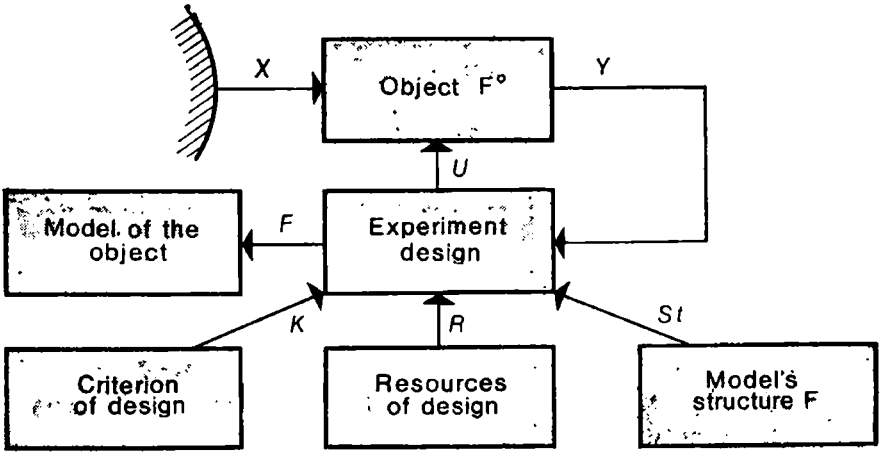

Fig. 5.1.1. Diagram of the process for designing experiments with the object

control are specific—to determine the parameters of the object's model, or, to be more precise, to produce the model of the controlled object (correction of the structure of the model is also possible). As it is clear from Fig. 5.1.1, the role of the controlling device is played by the procedure of experiment design which indicates how to change the input $\mathbf{U}$ in the process of the experiment. The experiment itself is unambiguously determined by the design $\widetilde{\mathbf{U}}$, a list of the states of the input $\mathbf{U}$ to be realized in the object. Thus, the design $\widetilde{\mathbf{U}}$ for the static object has the form

$$\widetilde{\mathbf{U}} = \langle \mathbf{U}_1, \ldots, \mathbf{U}_N \rangle, \tag{5.1.1}$$

where $\mathbf{U}_i$ is the i-th experiment of the design, i.e. the i-th state of the input:

$$\mathbf{U}_i = (u_{1i}, \ldots, u_{qi}), \quad i = 1, \ldots, N.$$

The design (5.1.1) is completely determined by the so-called *matrix of design*

$$\widetilde{\mathbf{U}} = \begin{bmatrix} \mathbf{U}_1 \\ \vdots \\ \mathbf{U}_N \end{bmatrix} = \begin{bmatrix} u_{11} & \ldots & u_{q1} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ u_{1N} & \ldots & u_{qN} \end{bmatrix}. \tag{5.1.2}$$

The design for the dynamic object is the vector function of time $\tilde{\mathbf{U}} = \mathbf{U}(t)$, i.e. the dependence of the input $\mathbf{U}$ on time.

The aims of designing an experiment, on the basis of which the experiment design $\tilde{\mathbf{U}}$ is produced, are usually determined by two factors: the resources and the criterion for designing.

The *design resources* R include the means P allocated for the experiment (time, materials, personnel, etc.), as well as the area (zone) of design Ω indicating the limits of change of the input $\mathbf{U}$ in the process of design:

$$\tilde{\mathbf{U}} \in \Omega. \tag{5.1.3}$$

Therefore

$$R = \langle P, \Omega \rangle. \tag{5.1.4}$$

For instance, it may be $P = N$, i.e. the volume of the design, or the allowed number of experiments with the object. The area Ω in this case is $a_i \leqslant u_i \leqslant b_i$ $(i = 1, \ldots, q)$, i.e. a hyperparallelepiped in the space of states of the input $\mathbf{U}$ whose values of the limits a_i and b_i $(i = 1, \ldots, q)$ should be specified.

The *design criterion* determines the efficiency of the design $\tilde{\mathbf{U}}$. The value of such a criterion naturally depends on the design $\tilde{\mathbf{U}}$ and the structure of the model $\mathbf{F}$ of the object:

$$K = K(\tilde{\mathbf{U}}, \mathbf{F}). \tag{5.1.5}$$

Usually a characteristic showing the accuracy of the estimates of the parameters $\mathbf{C}$ of the model is selected as such a criterion. For instance, this may be the sum, or the product of the dispersions of estimates of the determined parameters, or the maximal dispersion. Obviously, the minimization of such a criterion by means of the design $\tilde{\mathbf{U}}$ is desirable in the experiment design.

Sometimes there may arise a situation when it would be necessary to minimize not only one, but several (p) criteria. Then the following multicriterial problem should be solved:

$$K_i(\tilde{\mathbf{U}}, \mathbf{F}) \underset{\tilde{U} \in R}{\rightarrow} \min, \quad i = 1, \ldots, p. \tag{5.1.6}$$

It is necessary to make use of some additional information on the object in order to solve this problem (methods to solve such problems are given in the 'Supplement'). Various compromises between the criteria are the basis for the solution. Therefore, the designs synthesized while solving multicriterial problem (5.1.6) are called compromising.

Clearly, the aims of designing are determined proceeding from the specificity of the object of control and the objectives of its maintenance and control. We should not forget that the complex object we deal with exists and operates irrespectively of the objectives of the future control, and therefore any intervention into its operation before that control should be performed with great caution. Above all, it concerns selection of the area Ω of the design.

When selecting the dimensions of the area, an unavoidable conflict occurs between 'the master' of the object, i.e. the person responsible for its normal operation, and the designer of the control system. The former tends to minimize the size of the area Ω so that to keep the disturbance of the object normal functioning at its minimum, while the latter tries to maximize the size of the area Ω because then the efficiency of the experiment $\tilde{\mathbf{U}}$ is the greatest. Indeed, the more the states of the experimental inputs (5.1.1) differ, the more data on the object are obtained. The conflict is usually settled by means of a reasonable compromise between the two sides and this results in selecting concrete contours of the area of design Ω.

Consequently, to realize the aims of the design is to solve the following extremal problem:

$$K(\tilde{\mathbf{U}}, \mathbf{F}) \to \min_{\tilde{\mathbf{U}} \in R} \Rightarrow \tilde{\mathbf{U}}^*_K, \qquad (5.1.7)$$

i.e. varying the design $\tilde{\mathbf{U}}$ within the allocated resources R, we should minimize the given criterion at the structure of the model $\mathbf{F}$ specified earlier. The obtained design is $\tilde{\mathbf{U}}^*_K$; it is natural to call it the K-optimal design with the resources R for the object having the model $\mathbf{F}$. In case the said factors (the criterion, resources, or the object)

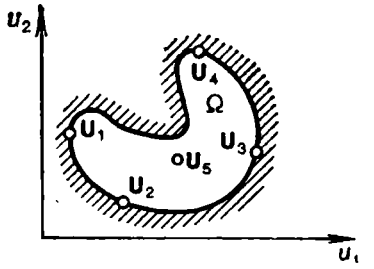

Fig. 5.1.2. The zone of design Ω in the space of controlled inputs {U}.
The points designate the experiments forming the design U*

changed, the optimal design $\tilde{\mathbf{U}}_K^*$ would also change.

An example of the optimal design $\tilde{\mathbf{U}}_K^*$ including $N = 5$ experiments for a static object (F is a function) with a specified structure is shown in Fig. 5.1.2. Each experiment is presented by a point in the space of the inputs {U}, and the design $\tilde{\mathbf{U}}_K^*$ is the set of five points belonging to the given area (zone) Ω of the design:

$$\tilde{\mathbf{U}}_K^* = \langle \mathbf{U}_1, \ldots, \mathbf{U}_5 \rangle, \ \mathbf{U}_i \in \Omega, \ i = 1, \ldots, 5;$$

the location of the points minimizes the given criterion K.

Let us note that the control U is very often of a discrete character. For instance, an i-th handle of the object control can be only in two positions: 'on' ($u_i = 1$) and 'off' ($u_i = 0$), and there are no intermediate positions. This specificity of the object is taken into account by a proper definition of the set Ω which consists, when discrete, of a finite number of elements. Thus, if an i-th parameter u_i may take on only k_i values, then the number of the elements of the set Ω is

$$k_1 k_2 \ldots k_q = \prod_{i=1}^{q} k_i.$$

For binary control ($k_i = 2$, $i = 1, \ldots, q$) there are 2^q elements in the set.

Consequently, the efficient solution of problem (5.1.6), i.e. the creation of optimal designs $\tilde{\mathbf{U}}^*$ with the admissible set of resources R, is the most important thing in experiment design.

Let us consider only static objects for whose models $\mathbf{F}$ is a function. The design $\widetilde{\mathbf{U}}$ in this case is a set of N points $\mathbf{U}_1, \ldots, \mathbf{U}_N$ in the space of varied parameters $\{\mathbf{U}\}$, all the points belonging to the zone of the design: $\mathbf{U}_i \in \Omega$ $(i = 1, \ldots, N)$.

The experiment design for the dynamic object (when $\mathbf{F}$ is an operator) is a function $\widetilde{\mathbf{U}} = \mathbf{U}(t) \in \Omega$ presenting a trajectory in the zone Ω of the design, the trajectory minimizing a selected criterion K which is a functional in this case*. The approaches to solve such problems are discussed in the next chapter.

Now let us consider methods to determine the criterion K.

5.2. Criteria for Experiment Design

Let us consider an experiment whose aim is the most reliable determination of the parameters $\mathbf{C} = (c_1, \ldots, c_k)$. But the concept of 'reliability' is too wide and blurred. Therefore, the determination of the criterion in each concrete case is definitely connected with the objectives $\mathbf{Z}^*$ of control, but also always depends on the characteristics of the accuracy of the sought parameters. Such characteristics are, for instance, the dispersion of the parameters $\mathbf{C}$

$$D_{ii} = M[(c_i - \bar{c}_i)^2], \quad i = 1, \ldots, k,$$

(where $\bar{c}_i$ is the mean value of the parameter c_i and M is the symbol of mathematical expectation) and the numbers

$$D_{ij} = M[(c_i - \bar{c}_i)(c_j - \bar{c}_j)],$$
$$i, j = 1, \ldots, k, \quad i \neq j$$

presenting the correlation of the parameters c_i and c_j. These characteristics jointly form the so-called dispersion

* Let us recall that a functional is the transformation of a function into a number, e.g.:

$$\int_a^b |\mathbf{U}(t)|^2 \, dt, \quad \max_{0 \leqslant t \leqslant T} |\mathbf{U}(t)| \text{ etc.}$$

matrix for the parameters C which is commonly defined as

$$D_C = \left\| \begin{matrix} D_{11} & \dots & D_{1k} \\ \dots & \dots & \dots \\ D_{k1} & \dots & D_{kk} \end{matrix} \right\|. \qquad (5.2.1)$$

Clearly, dispersion matrix (5.2.1) includes the complete set of correlation properties of the sought parameters C.

It is natural to ask: how to determine the dispersion matrix? In the general case, the problem is rather difficult. But for the models linear with respect to the parameters C, i.e. the models of the (3.4.3) kind, it is known that the dispersion matrix is the reverse of the Fisher information matrix, i.e.

$$D_C = \sigma^2 \Phi^{-1}. \qquad (5.2.2)$$

This means that each element of the dispersion matrix is

$$D_{ij} = \sigma^2 \Phi_{ji} / |\Phi|, \quad i, j = 1, \dots, k. \qquad (5.2.3)$$

Here Φ_{ji} is the algebraic complement to the element ψ_{ji} of the matrix Φ which is produced from Φ by means of crossing out the j-th row and the i-th column and multiplying it by $(-1)^{i+j}$; $|\Phi|$ is the determinant of the matrix Φ; σ^2 is the dispersion of random disturbance superimposed on the output of the object in compliance with structure (3.4.9).

Consequently, knowing the matrix Φ and using (5.2.3), it is easy to obtain the dispersion matrix D_C which makes it possible to produce any criteria required for the design.

The most common is the so-called D-criterion equal to the determinant of dispersion matrix (5.2.1), i.e. to the generalized dispersion:

$$K = |D_C|. \qquad (5.2.4)$$

The design minimizing this criterion is called D-optimal.

Generally speaking, any function of the elements of the dispersion matrix $K = \varphi(D_C)$ may be the criterion for the design. For instance, K may be the maximal dispersion ($K = \max_{i=1,\dots,q} D_{ii}$). The design minimizing this dispersion is called E-optimal. In case K is the sum

of the dispersion of the sought parameters ($K = \sum_{i=1}^{q} D_{ii}$), i.e. the trace of the dispersion matrix*, then the corresponding design is called A-optimal.

A special place among the numerous experiment designs is held by the orthogonal designs for which the Fisher information matrix Φ is diagonal. The orthogonal design makes the procedure of calculating the parameters $\mathbf{C}$ very simple:

$$c_i = \psi_{ii}/\eta_i, \quad i = 1, \ldots, k,$$

where ψ_{ii} and η_i are those defined in (4.2.11). But the main advantage of the orthogonal design is not that, but mutual non-correlation of the estimates of the sought parameters. Indeed, if the information matrix Φ is diagonal, then the dispersion matrix D is diagonal as well, from which it follows that all the correlation coefficients are equal to zero:

$$D_{ij} = 0, \quad i \neq j = 1, \ldots, k,$$

which means, as it was already mentioned, that the obtained estimates of the sought parameters $\mathbf{C}$ do not correlate. This is often very important for the synthesis of the model.

Now let us consider the most common D-optimal experiment design.

5.3. *D*-Optimal Design

It is easy to see that the synthesis of the D-optimal design may be reduced either to minimization of the determinant of dispersion matrix (5.2.1) or, considering (5.2.2), to maximization of the determinant of the information matrix Φ:

$$|\Phi(\tilde{\mathbf{U}})| \to \max_{\tilde{\mathbf{U}} \in \Omega} \Rightarrow \tilde{\mathbf{U}}^*_D. \tag{5.3.1}$$

The resulting design $\tilde{\mathbf{U}}^*_D$ is called D-optimal.

---

* Let us recall that the trace of a matrix is the sum of its diagonal elements.

Let us find a D-optimal design for a linear unidimensional object

$$y=c_1+c_2u. \tag{5.3.2}$$

The elements of the information matrix in this case may be determined according to (4.2.11), where $\varphi_1 = 1$, $\varphi_2 = u$. However, for the sake of clarity of representation, let us produce all the required calculations for the simple case under discussion ($n = q = 1$). In order to make use of the method of least squares while finding the parameters c_1 and c_2 of model (5.3.2), let us conditionally introduce initial information as the results of some experiments with the object:

$$I = \langle u_i, y_i \rangle, \quad (i = 1, \ldots, N), \tag{5.3.3}$$

where $\widetilde{U} = \langle u_1, \ldots, u_N \rangle$ is the sought experiment design.

The total discrepancy between the outputs of model (5.3.2) and object (5.3.3)

$$Q(c_1, c_2)=\sum_{i=1}^{N} [c_1+c_2u_i-y_i]^2.$$

Minimizing this expression by c_1 and c_2 in compliance with the least-squares method, we obtain a system of two equations

$$\frac{\partial Q}{\partial c_1}=2\sum_{i=1}^{N}(c_1+c_2u_i-y_i)=0,$$

$$\frac{\partial Q}{\partial c_2}=2\sum_{i=1}^{N}(c_1+c_2u_i-y_i)\,u_i=0.$$

The system may be conveniently transformed into

$$c_1+c_2\bar{u} \;=\frac{1}{N}\eta_1,$$

$$c_1\bar{u}+c_2\bar{u^2}=\frac{1}{N}\eta_2,$$

where

$$\bar{u}=\frac{1}{N}\sum_{i=1}^{N}u_i; \quad \bar{u^2}=\frac{1}{N}\sum_{i=1}^{N}u^2{}_i.$$

It follows from this that the determinant of the information matrix has the form

$$|\Phi(\widetilde{U})| = \begin{vmatrix} 1 & \bar{u} \\ \bar{u} & \bar{u^2} \end{vmatrix} N^2=[\bar{u^2}-(\bar{u})^2]\,N^2. \tag{5.3.4}$$

Now let us solve the problem to synthesize the D-optimal design by maximizing this expression for the elements of the

design $\widetilde{U}$, parameters u_i $(i = 1, \ldots, N)$. For the purpose, it is necessary to solve the problem of maximization

$$\overline{u^2} - (\overline{u})^2 \to \max_{u_i \in \Omega\ (i=1,\ldots,N)}$$

where Ω: $[a, b]$, i.e. the problem is to select N such numbers u_i within the segment $[a, b]$, i.e. $a \leqslant u_i \leqslant b$, that would maximize quadratic form (5.3.4).

The problem is solved easily by means of the following simple argumentation. Obviously, expression (5.3.4) is none other than the dispersion of the points of the design $\widetilde{U} = \langle u_1, \ldots, u_N \rangle$. To maximize the dispersion of N points within a segment, it is necessary that they should be located at the ends of the segment, a half of them at one end and the other half at the other.

Therefore, the D-optimal design $\widetilde{U}_D^* = \langle u_1^*, \ldots, u_N^* \rangle$ with an even N is as follows:

$$u^*_i = \begin{cases} a \text{ for } i \leqslant N/2, \\ b \text{ for } i > N/2. \end{cases}$$

if N is odd, there are two possible equivalent designs:

$$u^*_i = \begin{cases} a \text{ for } i \leqslant (N-1)/2, \\ b \text{ for } i > (N-1)/2 \end{cases}$$

or

$$u^*_i = \begin{cases} a \text{ for } i \leqslant (N+1)/2, \\ b \text{ for } i > (N+1)/2. \end{cases}$$

Let us note that in this case information matrix (5.3.4) is not diagonal, the design is not orthogonal, and therefore the obtained estimates of the parameters c_1 and c_2 are correlated. Indeed, their dispersion matrix is

$$D_C = \sigma^2 \Phi^{-1} = \sigma^2 \left\| \begin{matrix} \overline{u^2} & -\bar{u} \\ -\bar{u} & 1 \end{matrix} \right\| \frac{1}{N^2}. \qquad (5.3.5)$$

But this inconvenience is easy to negotiate. The design $\widetilde{U}^*$ can be made orthogonal by means of changing the reference system so that to obtain $\bar{u} = 0$ which, as it is clear from (5.3.5), makes the matrix Φ diagonal. To achieve this, it is sufficient to pass over to a centred reference system, i.e. to put the origin of the coordinates in the centre of the zone of design Ω making

$$u' = -\frac{a+b}{b-a} + \frac{2}{b-a} u,$$

for which $\bar{u}' = 0$, $(\overline{u'})^2 = 1$. Now the matrix has the required form

$$\Phi' = \left\| \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \right\| \frac{1}{N^2}.$$

The design in this case is

$$\tilde{U}^*_D = (\underbrace{-1, \ldots, -1}_{k}, \ldots, 1), \quad \underbrace{\qquad\qquad}_{N}$$

where

$$k = \begin{cases} N/2 & \text{for even } N, \\ (N+1)/2 & \text{for odd } N. \end{cases}$$

It is clear that the correlation properties of the sought parameters depend essentially on the selection of the reference system.

Applying similar argumentation to the D-optimal experiment design in the case of a linear multidimensional model

$$y = c_1 + c_2 u_1 + \ldots + c_{q+1} u_q$$

in the area Ω: $-1 \leqslant u_i \leqslant 1$ (a hypercube in the space $\{\mathbf{U}\}$), we shall get the D-optimal design $\tilde{\mathbf{U}}^*_D$ with coordinates $u_j^* = \pm 1$ $(j = 1, \ldots, q)$, i.e. the points $\mathbf{U}_i$ $(i = 1, \ldots, N)$ should be in the corners of the hypercube so that the design $\tilde{\mathbf{U}}$ would be orthogonal at that. The conditions of orthogonality are

$$\sum_{i=1}^{N} u_{ji} u_{li} = 0 \text{ at } i \neq l$$

i.e. the products of any two columns of the design matrix (5.1.2) are equal to zero. Such a matrix is called orthogonal with regard to columns. Examples of various designs for $n = 2$, when Ω is a square, are shown in

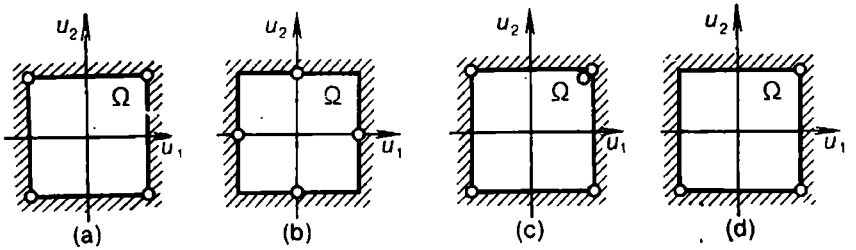

Fig. 5.3.1. **Examples of designs:**
(*a*)—*D*-optimal orthogonal (N=4); (*b*)—orthogonal, but not *D*-optimal (N=4); (*c*)—*D*-optimal non-orthogonal (N=5); (*d*)—the same for N=3

Fig. 5.3.1. Their design matrices are correspondingly

$$\widetilde{U}^*_D = \widetilde{U}^0 = \begin{Vmatrix} 1 & 1 \\ -1 & 1 \\ -1 & -1 \\ 1 & -1 \end{Vmatrix}; \quad \widetilde{U}^0 = \begin{Vmatrix} 0 & 1 \\ -1 & 0 \\ 0 & -1 \\ 1 & 0 \end{Vmatrix};$$

$$\widetilde{U}^*_D = \begin{Vmatrix} 1 & 1 \\ 1 & 1 \\ -1 & 1 \\ -1 & -1 \\ 1 & -1 \end{Vmatrix}; \quad \widetilde{U}^*_D = \begin{Vmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \end{Vmatrix}.$$

Now let us consider an example of a quadratic dependence **F** in the simplest unidimensional case

$$y = c_1 + c_2 u + c_3 u^2,$$

where functions (4.2.8) of decomposition are: $\varphi_1(u) = 1$; $\varphi_2(u) = u$; $\varphi_3(u) = u^2$. It is easy to show that the Fisher information matrix (4.2.12) acquires the form

$$\Phi = \begin{bmatrix} 1 & \overline{u} & \overline{u^2} \\ \overline{u} & \overline{u^2} & \overline{u^3} \\ \overline{u^2} & \overline{u^3} & \overline{u^4} \end{bmatrix} N^3 \qquad (5.3.6)$$

where

$$\overline{u^j} = \frac{1}{N} \sum_{i=1}^{N} u^j{}_i, \quad j = 1, 2, 3, 4.$$

It is interesting that this is a case when (with the given variables) an orthogonal design does not exist, since at $\overline{u} = \overline{u^2} = \overline{u^3} = 0$, the determinant $|\Phi|$ becomes equal to zero, which excludes the possibility to determine **C**. Therefore, to make the synthesis of the design easier, it is only natural to demand

$$\overline{u} = \overline{u^3} = 0, \qquad (5.3.7)$$

i.e. the design should be symmetric with regard to the centre of the zone of design Ω.

Suppose Ω is $-1 \leqslant u \leqslant 1$, then the determinant of the Fisher information matrix

$$|\Phi| = \begin{vmatrix} 1 & 0 & \overline{u^2} \\ 0 & \overline{u^2} & 0 \\ \overline{u^2} & 0 & \overline{u^4} \end{vmatrix} N^3 = \{\overline{u^2}\,[\overline{u^4} - (\overline{u^2})^2]\}\, N^3.$$

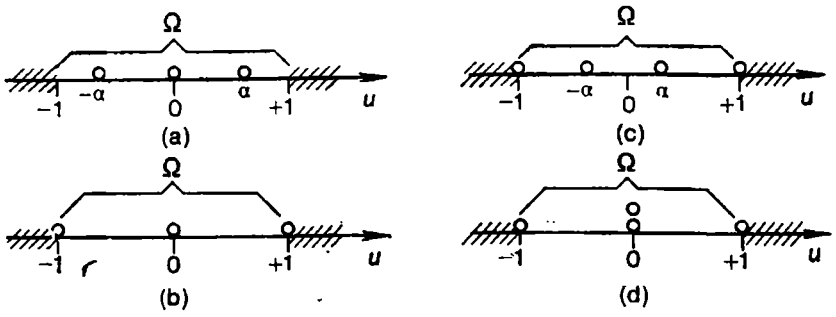

Fig. 5.3.2. To finding the D-optimal designs on the segment for quadratic dependence F:
(a, c)—non-optimal design; (b, d)—optimal design

Now let us produce the optimal designs with various volumes N, i.e. vary the number of points in the design.

Suppose $N = 3$. A symmetrical design in this case is determined by one parameter α and has the form

$$\widetilde{U} = \langle -\alpha, 0, \alpha \rangle$$

(Fig.5.3.2a). The second and fourth moments are

$$\overline{u^2} = \frac{2}{3}\alpha^2; \quad \overline{u^4} = \frac{2}{3}\alpha^4$$

and the problem to synthesize an optimal design is reduced to maximizing the expression

$$|\Phi| = \frac{4}{27}\alpha^6 \to \max_{-1 \leqslant \alpha \leqslant 1} \Rightarrow \alpha^* = 1$$

whose solution is obvious. Therefore, the D-optimal design in this case has the form $\widetilde{U}_D^* = \langle -1, 0, +1 \rangle$ (see Fig. 5.3.2b).

Suppose $N = 4$. Then, fixing two obvious points of the design at the ends ($u = \pm 1$), let us consider a symmetrical design with one unknown parameter $\alpha \geqslant 0$: $\widetilde{U} = (-1, -\alpha, \alpha, 1)$ (see Fig. 5.3.2c) for which

$$\overline{u^2} = \frac{1}{2}(1+\alpha^2); \quad \overline{u^4} = \frac{1}{2}(1+\alpha^4)$$

and the determinant is

$$|\Phi| = \frac{1}{8}(1-\alpha^2)(1-\alpha^4).$$

Its maximum is obvious at $\alpha = 0$ which yields the D-optimal design

$$\tilde{U}_D^* = \langle -1,\ 0,\ 0, +1 \rangle$$

(see Fig. 5.3.2*d*) for which $|\Phi| = \frac{1}{8}$. Note that this design differs from the one produced previously ($N = 3$) by one more zero point.

Now let us consider a non-symmetrical design $u_1 = -1$, $u_2 = 0$, $u_3 = u_4 = 1$. Here $|\Phi| = \frac{1}{8}$, i.e. this design

$$\tilde{U}^*{}_D = \langle -1,\ 0,\ +1,\ +1 \rangle$$

is also D-optimal.

The analysis shows that for the quadratic object the elements of the D-optimal design at any N are grouped in the three points $u = -1,\ 0,\ +1$. Therefore, when synthesizing a design for $N > 4$, we should manipulate by these points only.

In the general case, when there are k points $u = -1$, q points $u = 0$, and m points $u = +1$ ($k + q + m = N$, $k > m$), then $\bar{u} = \overline{u^3} = (k - m)/N$, $\overline{u^2} = \overline{u^4} = (k + m)/N$, and we can get from (5.3.6)

$$|\Phi| = 4\,kqmN^2.$$

Obviously, to maximize this expression, it is sufficient to have $k = q = m = N/3$—naturally, if N is a multiple of 3. Otherwise k, q, and m should differ by no more than unity. Thus, for instance, the designs *a* and *b* in Fig. 5.3.3 are optimal, while *c* and *d* are non-optimal.

It follows that an optimal design with the volume of $(N + 1)$ may be obtained from an optimal design with the volume of N by adding an experiment so that the new design should be as uniform as possible.

This rule can be generalized over other designs. Thus, for $n = 2$ the D-optimal design on the square $-1 \leqslant u_i \leqslant 1$ ($i = 1, 2$) for the quadratic structure is naturally based on the points shown in Fig. 5.3.4*a*. These points form a D-optimal design for $N = 9$*. The optimal design for $N = 8$ may be produced from the previous one by means of rejecting any of the four points (*1*, *2*, *3*, *4*) on the axes of the square Ω (see, for instance, Fig. 5.3.4*b*). For $N = 7$, we should reject two of those points (Fig. 5.3.4*c*), for $N = 6$, any three points on the axes except

* Strictly speaking, this design is not optimal [4], but it differs from the optimal one quite insignificantly.

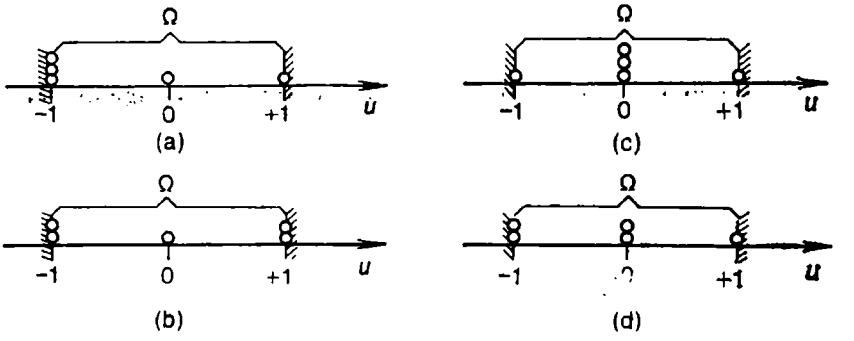

Fig. 5.3.3. To finding the D-optimal design at $N = 5$; (a, b)—optimal designs; (c, d)—non-optimal designs

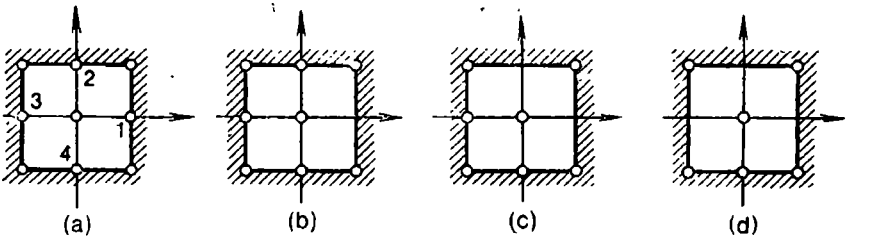

Fig. 5.3.4. D-optimal designs for the quadratic model ($n = 2$): (a) $N=9$; (b) $N=8$; (c) $N=7$; (d) $N=6$

for the central one (Fig. 5.3.4d). The number of points for the optimal design may be increased by adding those basic points (see Fig. 5.3.4a) which give the maximal increment to the determinant $|\Phi|$.

This rule is valid for any other design of any dimension, which allows for easy recurrent construction of the optimal designs by means of consecutive adding the basic points to the optimal design for $(N - 1)$ synthesized before.

5.4. Step-by-Step Design of Experiments

Very commonly the volume N of the design is not predetermined and depends on what it may give for the synthesis of the model, as well as on the persistence of the experimenter. Therefore, it is natural not to stipulate the volume N beforehand, but determine it in the process of the experiments with the object. This is provided by means

of a step-by-step experiment design when the decision on the next experiment is made only after the previous one has been completed. Besides, in the process of the step-by-step experiment design, the experimenter has a possibility to change the structure of the model, i.e. to pass over from one description of the object to another, change the zone of the design and the optimum criterion.

The step-by-step design is supported by the obvious assumption that the data obtained in an experiment can be used not only for the correction of the model's parameters (as it was done in the identification processes), but also for making other decisions (for instance, to change the model's structure, stop the experiments, change the criterion or the zone of design, etc.).

Suppose $(i - 1)$-th stage of design resulted in $(i - 1)$ experiments in the points $\widetilde{\mathbf{U}}_{i-1} = \langle \mathbf{U}_1, \ldots, \mathbf{U}_{i-1} \rangle$, i.e. the design $\widetilde{\mathbf{U}}_{i-1}$ has been carried out. The obtained data $y_1, \ldots, y_{i-1}$ made it possible to synthesize the model

$$y = f_{i-1}(\mathbf{U}, \mathbf{C}_{i-1}) \qquad (5.4.1)$$

(for instance, by the least-squares or some other method), where f_{i-1} is the structure of the model which after the $(i - 1)$-th stage reflects its specificity and best corresponds to the objectives of control in the opinion of the experimenter; $\mathbf{C}_{i-1}$ are the values of the model's parameters at the $(i-1)$-th stage of the design. Having such a model, the experimenter should decide first whether or not the model meets his requirements. In case it does, the model synthesis has come to an end; if it does not, the experimenter should make an estimate whether the more accurate model to be produced would cover the expenditures for the realization of the i-th stage of the experiments. In case it would not, the experiments stop; if it would, the i-th stage of the experimenting should be started. It is necessary to determine again:

—a new structure f_i of the model, if the experimenter decides to change it;

—a new zone Ω_i for the experiment design, if there is a need and possibility for that;

—a new criterion K_i for the design, if it is necessary to change it.

In case there are no changes, $f_i = f_{i-1}$, $\Omega_i = \Omega_{i-1}$ $K_i = K_{i-1}$.

It is known that the numerical value of the criterion K depends on the structure of the model, i.e. on the selected system of the functions $\{\varphi_j(\mathbf{U})\}$ $(j = 1, \ldots, k)$ representing the model

$$y = f(\mathbf{U}, \mathbf{C}) = \sum_{j=1}^{k} c_j \varphi_j(\mathbf{U}) \qquad (5.4.2$$

and on the experiment design $\widetilde{\mathbf{U}}$(5.1.3). In the step-by-step design, only one i-th experiment is designed at the i-th stage; it is carried out in the state $\mathbf{U}_i$, i.e. the design is

$$\widetilde{\mathbf{U}}_i = \langle \widetilde{\mathbf{U}}_{i-1}, \mathbf{U}_i \rangle, \qquad (5.4.3$$

where $\widetilde{\mathbf{U}}_{i-1}$ is the design already realized by the i-th stage.

The state for the i-th experiment, i.e. $\mathbf{U}_i$, is determined as the result of solving the following obvious optimization problem:

$$K_i(f_i, \widetilde{\mathbf{U}}_{i-1}, \mathbf{U}_i) \to \min_{\mathbf{U}_i \in \Omega_i} \Rightarrow \mathbf{U}^*_i. \qquad (5.4.4$$

This problem is very complicated and it should be solved by one of the optimum seeking methods considered in the 'Supplement'. Its solution is $\mathbf{U}_i^*$.

The state $\mathbf{U}_i^*$ is to be reached to carry out the experiment, then the output y_i of the object is determined, etc. If necessary, the model's structure, the zone of design and the criterion are determined again (or kept unchanged) at the next $(i + 1)$-th stage of design.

Naturally, it is possible to design not only one, but several experiments, i.e. $\mathbf{U}_i$ in expression (5.4.4) may be substituted by $\mathbf{U}_i, \ldots, \mathbf{U}_{i+k} \in \Omega_i$, by means of which the criterion K_i should be minimized.

Thus, all the preceding experiments $\widetilde{\mathbf{U}}_{i-1}$ are taken into account and changes are introduced to make more accurate or substitute the criteria, structures, etc.

The consecutive design is therefore a very flexible instrument; promptly changing the profile of the experiments, the experimenter has a possibility to obtain the required model of the object in the long run.

5.5. Discriminating Designs

The designs which allow for the selection of one among several hypotheses on the structure of the connection between the input and the output of the object, i.e. on the structure of the model, are usually called discriminating designs. Suppose there are p competing structures (models)

$$y = f_j(\mathbf{U}, \mathbf{C}_j), \quad j = 1, \ldots, p, \tag{5.5.1}$$

where $\mathbf{C}_j$ is the vector of the parameters of the j-th model:

$$\mathbf{C}_j = (c_{1j}, \ldots, c_{k_j j}).$$

Generally, the number k_j of the parameters may change from one model to another.

Suppose there is a design consisting of N experiments $\tilde{\mathbf{U}} = \langle \mathbf{U}_1, \ldots, \mathbf{U}_N \rangle$, and there are observations $y_1, \ldots, y_N$ of the output of the object while the design is realized. Then the selection of the best structure of the model is based on the determination of the resulting discrepancy for each of the models

$$Q_j = \min_{\mathbf{C}_j} \sum_{i=1}^{N} [f_j(\mathbf{U}_i, \mathbf{C}_j) - y_i]^2, \quad j = 1, \ldots, p. \tag{5.5.2}$$

The structure providing the least discrepancy is selected. That means that the best structure is f_l if

$$Q_l = \min_{j=1, \ldots, p} Q_j. \tag{5.5.3}$$

Let us call optimal discriminating the design $\tilde{\mathbf{U}}^*$ which allows for the best realization of this procedure. Let us note that the decision to select a structure with the minimal discrepancy (5.5.3) may prove to be wrong since the results of the discriminating experiments (as well as of any real ones) are of a stochastic nature. Therefore, the problem to synthesize the optimal discriminating design is reduced to making the probability of the error minimal, i.e. the decision to select the best structure has to have the greatest reliability. Let us formalize it.

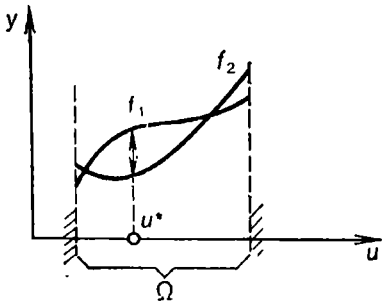

Fig. 5.5.1. To discrimination between two unidimensional models

Let β be the probability for the wrong decision, i.e.

$$\beta = \text{Prob}\,[f^* \neq f_l | Q_{\min} = Q_l], \qquad (5.5.4)$$

where f^* is the best structure, $Q_{\min}$ is the least discrepancy (5.5.2). This probability depends on the discriminating design $\tilde{\mathbf{U}}$: $\beta = \beta(\tilde{\mathbf{U}})$. Then the design that minimizes this probability is natural to be called the optimal discriminating design $\tilde{\mathbf{U}}^*$. To obtain it, it is necessary to solve the problem

$$\beta(\tilde{\mathbf{U}}) \to \min_{\tilde{\mathbf{U}} \in \Omega} \Rightarrow \tilde{\mathbf{U}}^*. \qquad (5.5.5)$$

Let us discuss the methods to solve it, considering, for simplicity, the case when the values of the parameter vectors $\mathbf{C}_j$ $(j = 1, \ldots, p)$ for each structure are known beforehand and therefore may not be taken into account i.e. $f_j(\mathbf{U}, \mathbf{C}_i) = f_j(\mathbf{U})$. Now the selection of the discriminating design may be reduced to the determination of N points of the experiments $\mathbf{U}_i$ $(i = 1, \ldots, N)$ and the calculation of total discrepancies in them by formula (5.5.2).

Suppose in a particular case $(p=2)$ there are two models f_1 and f_2 (Fig. 5.5.1). Obviously, the discriminating design at the point where the two models differ most should be considered optimal. It is clear that the selection of that design is equivalent to solving problem (5.5.5 (further on this equivalence is going to be proved strictly). Therefore, in order to find the only point of the optimal discriminating design, it is sufficient to solve the

problem

$$[f_1(\mathbf{U}) - f_2(\mathbf{U})]^2 \to \max_{\mathbf{U}\in\Omega} \in \mathbf{U}^*. \qquad (5.5.6)$$

Its solution, the point $\mathbf{U}^*$, represents the sought design: all N experiments should be located in this point, i.e. $\mathbf{U}_i = \mathbf{U}^*$ $(i = 1, \ldots, N)$.

To make it definite, let us consider the case when one of the competing structures is real (suppose it is f_1, i.e. $f^o = f_1$). The output of the object in the experiment has the form

$$y = f^o(\mathbf{U}) = f_1(\mathbf{U}) + \varepsilon, \qquad (5.5.7)$$

where ε is the random disturbance presupposed to be centred and normal, i.e. with normal distribution, zero expectation, and the dispersion σ^2. Therefore, the results of all N experiments y_i $(i=1, \ldots, N)$ will be grouped around the value f_1 $(\mathbf{U})$ in compliance with the distribution of the disturbance ε. Let us find the values of the discrepancies Q_j $(j = 1, 2)$. Taking into account (5.5.7), we get

$$Q_1 = \sum_{i=1}^{N} \varepsilon^2_i, \qquad (5.5.8)$$

$$Q_2 = \sum_{i=1}^{N} [f_2(\mathbf{U}) - f_1(\mathbf{U}) - \varepsilon_i]^2.$$

Let $\Delta(\mathbf{U})$ be the difference between the outputs of the models at the point $\mathbf{U}$:

$$\Delta(\mathbf{U}) = f_2(\mathbf{U}) - f_1(\mathbf{U}). \qquad (5.5.9)$$

Then after obvious transformations, we get

$$Q_2 = N\Delta^2(\mathbf{U}) - 2\Delta(\mathbf{U}) \sum_{i=1}^{N} \varepsilon_i + \sum_{i=1}^{N} \varepsilon^2_i.$$

It is easy to note that the sign of the difference

$$\delta(\mathbf{U}) = Q_2(\mathbf{U}) - Q_1(\mathbf{U}) = N\Delta^2(\mathbf{U}) - 2\Delta(\mathbf{U}) \sum_{i=1}^{N} \varepsilon_i$$

determines the selection of one of the models. Indeed, if the difference is positive, we should choose f_1, if negative,

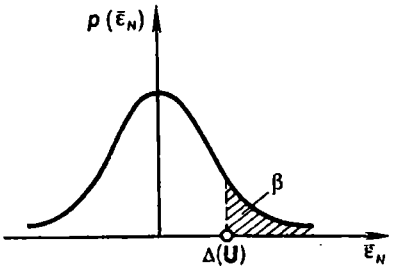

Fig. 5.5.2. To finding the probability of making the wrong decision in discriminating between two models

f_2 (let us remind that the figures Q_i are non-negative by definition). Consequently,

$$f_l = \begin{cases} f_1 \text{ when } \delta(U) \geq 0 \\ f_2 \text{ when } \delta(U) < 0. \end{cases} \tag{5.5.10}$$

This is the rule to solve the problem (5.5.6), and, therefore, (5.5.5).

Let us find the probability of the error, i.e. the probability to make the wrong decision, at $\Delta(U) > 0$:

$$\beta = \text{Prob}\left[\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i > \frac{1}{2}\Delta(U)\right], \tag{5.5.11}$$

i.e. the probability that the random value

$$\bar{\varepsilon}_N = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i$$

will exceed $\frac{1}{2}\Delta(U)$. Knowing the stochastic properties of $\bar{\varepsilon}_N$:

$$M(\bar{\varepsilon}_N) = 0;\ D(\bar{\varepsilon}_N) = \sigma^2/N,$$

it is easy to find the sought probability. It is equal to the cross-hatched area under the distribution curve for the random value $\bar{\varepsilon}_N$ (Fig. 5.5.2). For the normally distributed $\bar{\varepsilon}_N$:

$$\beta = \frac{1}{2}\left[1 - \Phi\left(\frac{\Delta(U)\sqrt{N}}{2\sqrt{2}\sigma}\right)\right], \tag{5.5.12}$$

where $\Phi(\cdot)$ is the Laplace function which increases monotonically. This property of the function $\Phi(\cdot)$ determines

the selection of the sought design which, in order to minimize the probability of the error, should maximize the value of Δ (U), i.e. solve problem (5.5.6). Substituting the obtained point $\mathbf{U}^*$ into expression (5.5.12), we get the error probability for the optimal design. On the other hand, the error probability for such a solution may be decreased by increasing the number N of the experiments in the point $\mathbf{U}^*$. Thus, it is possible to design experiments providing the selection of the model with any degree of reliability given beforehand.

But in real problems, the parameters $\mathbf{C}_j$ ($j = 1, \ldots, p$) of models (5.5.1) are unknown, and, therefore, it is necessary to estimate the parameters simultaneously with the discrimination. This may be done by means of a successive experiment design in the following way.

Suppose for simplicity that $p = 2$, i.e. the problem is again to discriminate between two models $f_1(\mathbf{U}, \mathbf{C}_1)$ and $f_2(\mathbf{U}, \mathbf{C}_2)$, where the parameter vectors $\mathbf{C}_1$ and $\mathbf{C}_2$ are unknown. At the $(i+1)$-th stage of the discrimination, we have concrete values of the parameters $\mathbf{C}_{1i}$ and $\mathbf{C}_{2i}$ for which the problem to synthesize the discriminating design should be solved

$$|f_1(\mathbf{U}, \mathbf{C}_{1i}) - f_2(\mathbf{U}, \mathbf{C}_{2i})| \to \max_{\mathbf{U} \in \Omega} \Rightarrow \mathbf{U}_{i+1}. \qquad (5.5.13)$$

Finding in the experiment the value of the object output

$$y_{i+1} = f^o(\mathbf{U}_{i+1}),$$

we can correct the parameters by means of the familiar method of adaptive identification (see § 4.2)

$$\mathbf{C}_{j,i+1} = \mathbf{C}_{j,i} + \Delta\mathbf{C}_{j,i+1}, \quad j = 1, 2, \qquad (5.5.14)$$

where

$$\Delta\mathbf{C}_{j,i+1} = -2\alpha_{i+1}q_{i+1}(\mathbf{C}_{j,i})\nabla_C f(\mathbf{U}_{i+1}, C_{j,i}).$$

The expressions included into this formula were determined above in (4.2.14) and (4.2.16). The obtained new values of the parameters make it possible to determine similarly the new point of discrimination $\mathbf{U}_{i+2}$, and so on.

The decision to select a model is made when there is an assurance that the parameters of the models have been

determined correctly. Now if

$$\sum_{i=1}^{N} [f_1(\mathbf{U}_i, \mathbf{C}_{1N}) - y_i]^2 < \sum_{i=1}^{N} [f_2(\mathbf{U}_i, \mathbf{C}_{2N}) - y_i]^2,$$

the first model is selected, and if the inequality is reverse, the second. The probability to make the wrong decision may be estimated by the approximate formula

$$\beta \approx 1 - \prod_{i=1}^{N} (1 - \beta_i), \qquad (5.5.15)$$

where

$$\beta_i = \frac{1}{2}\left[1 - \Phi\left(\frac{\Delta_i}{2\sqrt{2}\sigma}\right)\right],$$

$$\Delta_i = |f_1(\mathbf{U}_i - \mathbf{C}_{1i}) - f_2(\mathbf{U}_i, \mathbf{C}_{2i})|.$$

The disargeement between formulas (5.5.12) and (5.5.15) at $\mathbf{U}_i = \mathbf{U}$ is due to the approximate nature of estimate (5.5.15). They coincide at $N = 1$.

The problems to design discriminating experiments are of dual character. Two goals are to be reached at the same time: to determine the parameters of the models and to select the best. The contradiction in these requirements is obvious. Indeed, efficient determination of the parameters demands to set the points of the design on the margin of the area of design, while discrimination requires them where the models differ most. A compromise is needed, one of which has been presented above.

Consequently, the stages of the model's synthesis (the stages *3-5* in the diagram of Fig. I.4.1) terminate in the development of the model F of the object of control. Such a model is necessary for the synthesis of any complex object.

Comments on References to Chapter 5

Experiment design is a part of applied mathematical statistics and, therefore, it is not easily popularized.

To make the first acquaintance with the basic ideas of the experiment design, one can be recommended a very interesting and artlufly presented book [1], as well as less popular but very rich [2]. The term 'experiment design' has two meanings. First, these are the methods to carry out experiments with the real object to determine its model (this is the sense in which this term is used in this

book). But there is also another meaning to it: the design of extremal experiments, i.e. experimental determination of the optimum conditions for the object functioning. As a matter of fact, the latter trend in experiment design is a part of the theory of control for the objects of the 'black box' type, i.e. the objects which have no mathematical description. One can introduce oneself into this important for the practice trend by reading a not very difficult and highly informative book [3].

But the problem of experimental determination of unknown parameters of the object is to be solved in another manner when the structure of the object model is known. This kind of experiment design is tackled with in a number of very informative but not quite easy books (it is necessary to know the theory of probability and linear algebra). Let us note among them the most comprehensive [4-6], as well as [7-8].

The total of more than 100 various criteria for the design have been proposed and used (according to publications). They clearly fall into two classes. The first one includes the criteria determined on the dispersion matrix. The values of these criteria can be calculated without experiments. Some of them were discussed in this chapter. The second class includes the criteria involving the value of the discrepancy between the outputs of the object and its model, i.e. the adequacy of the model. To determine the values of these criteria, it is necessary to have the results of the experiments with the object.

References

1. Khurgin, Ya.I. *Yes, No, or Maybe* ... Nauka, Moscow, 1977 (in Russian).
2. Nalimov, V.V. *The Theory of the Experiment.* Nauka, Moscow, 1971 (in Russian).
3. Adler, Yu.P., Markova, E.V., Granovskii, Yu.V. *The Design of Experiments to Find Optimal Conditions.* Mir Publishers, Moscow, 1975.
4. Fyodorov, V.V. *The Theory of the Optimal Experiment: the Design of Regressive Experiments.* Nauka, Moscow, 1971 (in Russian).
5. Krug, G.K., Sosulin, Yu.A., Fatuev, V.A. *Design of the Experiment in the Problems of Identification and Extrapolation.* Nauka, Moscow, 1977 (in Russian).
6. Denisov, V.I. *Software for the System 'Computer-Experimenter': the Regressive and Dispersion Analyses.* Nauka, Moscow, 1977 (in Russian).
7. Himmelblau, D.M. *Process Analysis by Statistical Methods.* Wiley, Inc., New York, 1970.
8. Statistische Versuchsplanung und Auswertung in der Stoffwirtschaft. VEB Deutscher Verlag für Grundstoffindustrie, Leipzig, 1974.

Chapter 6 Synthesis of Control (Decision-Making)

To realize this stage, it is necessary to have the object model $\mathbf{F}(\mathbf{X}, \mathbf{U})$, the data on the state of the medium $\mathbf{X}$, and the objective of control $\mathbf{Z}^*$. The procedure to determine the control $\mathbf{U}^*$ that solves the proposed problem in the developing situation is called the algorithm of control. However, before considering this algorithm, let us formulate the problem.

6.1. Formulation of the Problem

The situation produced in the process of control is characterized by a three

$$\langle \mathbf{X}, \mathbf{Y}, \mathbf{Z}^* \rangle. \tag{6.1.1}$$

By means of control, it is possible to change only the state $\mathbf{Y}$ of the object whose model

$$\mathbf{Y} = \mathbf{F}(\mathbf{X}, \mathbf{U}) \tag{6.1.2}$$

is at our disposal. The achievement of the objectives of control $\mathbf{Z}^*$, as it was shown in Chapter 1, is equivalent to the realization of conditions (1.3.8). Considering that $\mathbf{S} = \langle \mathbf{X}, \mathbf{Y} \rangle$ (see (1.3.12)), they may be reduced to the following objective relationships:

$$\mathbf{Z}^*: \begin{cases} \psi_i(\mathbf{X}, \mathbf{Y}) = a_i & (i = 1, \ldots, k_1), \\ \psi_{k_1+j}(\mathbf{X}, \mathbf{Y}) \geqslant b_j & (j = 1, \ldots, k_2), \\ \psi_{k_1+k_2+l}(\mathbf{X}, \mathbf{Y}) \to \min & (l = 1, \ldots, k_3), \end{cases} \tag{6.1.3}$$

where the functions ψ and the numbers a_i and b_j should be fixed at the stage of the determination of the control objectives (see Chapter 1). Let us write these objective

conditions in their canonical vector form:

$$Z^*: \begin{cases} \mathbf{H}(\mathbf{X}, \mathbf{Y}) \geqslant 0, \\ \mathbf{G}(\mathbf{X}, \mathbf{Y}) = 0, \\ \mathbf{Q}(\mathbf{X}, \mathbf{Y}) \to \min, \end{cases} \tag{6.1.4}$$

where

$$\begin{aligned} \mathbf{G}(\mathbf{X}, \mathbf{Y}) &= (g_1(\mathbf{X}, \mathbf{Y}), \ldots, g_{k_1}(\mathbf{X}, \mathbf{Y})); \\ \mathbf{H}(\mathbf{X}, \mathbf{Y}) &= (h_1(\mathbf{X}, \mathbf{Y}), \ldots, h_{k_2}(\mathbf{X}, \mathbf{Y})); \\ \mathbf{Q}(\mathbf{X}, \mathbf{Y}) &= (q_1(\mathbf{X}, \mathbf{Y}), \ldots, q_{k_3}(\mathbf{X}, \mathbf{Y})); \end{aligned} \tag{6.1.5}$$

while

$$\begin{aligned} g_i(\mathbf{X}, \mathbf{Y}) &= \psi_i(\mathbf{X}, \mathbf{Y}) - a_i \quad (i = 1, \ldots, k_1) \\ h_j(\mathbf{X}, \mathbf{Y}) &= \psi_{k_1+j}(\mathbf{X}, \mathbf{Y}) - b_j \quad (j = 1, \ldots k_2) \\ q_l(\mathbf{X}, \mathbf{Y}) &= \psi_{k_1+k_2+l}(\mathbf{X}, \mathbf{Y}) \quad (l = 1, \ldots, k_3). \end{aligned}$$

The introduced functions may be called for convenience as follows: $\mathbf{G}(.,.)$—the function of the constraints of the type of equalities, $\mathbf{H}(.,.)$—the function of the constraints of the type of inequalities, $\mathbf{Q}(.,.)$—the quality function.

It is possible to realize functions (6.1.4) only by means of a proper change in the state $\mathbf{Y}$ of the object of control. This means that they may be realized only by a proper selection of the control $\mathbf{U}$ (the state $\mathbf{X}$ of the medium changes regardless of our efforts). These obvious considerations lead to the following extremum problem:

$$\mathbf{Q}(\mathbf{X}, \mathbf{Y}) \to \min_{\mathbf{U} \in \Omega} \Rightarrow \mathbf{U}^*, \tag{6.1.6}$$

where

$$\Omega: \begin{cases} \mathbf{H}(\mathbf{X}, \mathbf{Y}) \geqslant 0, \\ \mathbf{G}(\mathbf{X}, \mathbf{Y}) = 0, \\ \mathbf{Y} = \mathbf{F}(\mathbf{X}, \mathbf{U}), \end{cases}$$

whose solution $\mathbf{U}^*$ is the optimal control. The model $\mathbf{F}$ (6.1.2) has been used here and the resource allocated for the control has been taken into account—the latter is expressed in the form of proper equalities and inequalities within Ω.

Let us consider problem (6.1.6) in more detail. It requires to minimize $\mathbf{Q}$ by means of selecting the proper control $\mathbf{U}$ which has, in addition, to satisfy the condi-

tions Ω. There is an impression that we ranked, as it were, our objectives, considering the main ones to be the extremal objectives **Q** while putting the non-extremal ones into constraints Ω. But that kind of ranking is only seeming. As a matter of fact, the main thing is to realize the constraints Ω, and the extremal objectives **Q** come second: they are to be reached only if the non-extremal objectives, i.e. the constraints Ω, are realized.

Consequently, the problem to synthesize the control (6.1.6) is characterized by a pair $\langle \mathbf{Q}, \Omega \rangle$, i.e. by the extremal and non-extremal objectives, the achievement of the latter having to precede the achievement of the former.

This natural requirement is widely applied in our daily life. Thus, we are very careful not to violate the constraints Ω. These are, for instance, the traffic regulations, the moral and criminal codes, the rules of communal life, fire safety, etc. And it is only when these natural restrictions are observed that we reach our extremal objectives **Q** (success and fun, happiness, praise, and so on). The one who is too indulged in reaching the extremal objectives **Q** to the detriment of the constraints Ω risks to trespass against the laws, both moral and juridical.

Problem (6.1.6) may be conveniently transformed and put down as

$$\mathbf{Q}(\mathbf{X}, \mathbf{F}(\mathbf{X}, \mathbf{U})) \to \min_{\mathbf{U} \in \Omega}, \qquad (6.1.7)$$

where

$$\Omega: \begin{cases} \mathbf{G}(\mathbf{X}, \mathbf{F}(\mathbf{X}, \mathbf{U})) = 0, \\ \mathbf{H}(\mathbf{X}, \mathbf{F}(\mathbf{X}, \mathbf{U})) \geqslant 0. \end{cases}$$

Depending on the form of the model **F**, i.e. on whether **F** is a function or an operator, we get different problems to be solved by different methods.

Let us consider first the static object of control whose model **F** is a function. The problem to synthesize control in this case is to minimize the k_3 functions $q_l(\mathbf{X}, \mathbf{U})$ $(l = 1, \ldots, k_3)$ by means of varying the q parameters $u_1, \ldots, u_q$ meeting the constraints Ω superimposed on **U**. The constraints include the requirements for the given functions of the control **U** to be non-negative or equal to zero. The state **X** of the medium is naturally supposed to be known.

An example is the problem of planning industrial production. Here U is the production plan; X is the supply of raw materials; G is the requirements for the assortment of the products, the need to observe the industrial safety regulations, etc.; H is the requirements to the quality of the products (no worse than ...), various production standards (from... to ...) and restrictions (not more than ...). The extremal objectives Q in planning the production may be, for instance, labour productivity, cost of production, the total of its output, and the deviation from the planned target (this should be minimized). Model F of production here consists of considerations allowing determination of the state Y of production at some known state X of the medium and production target U. These considerations usually are of simulation character (see § 3.5).

Consequently, the problem to control static objects is characterized by that the sought control is a set of the parameters $u_1, \ldots, u_q$, while Q, G, and H are the vector functions of control U. Such problems are known as the problems of *mathematical programming*.

Now let us consider the dynamic object whose model F is an operator. The control U in this case is a vector function of time $\mathbf{U}(t)$; q_l, h_i and g_j are functionals, and problem (6.1.7) is called *variational*.

An example is a problem for a taxi-driver who has to pass from one non-regulated cross-road to another in a minimal time (non-regulation means a stop at every cross-road). The equation for the movement of the taxi, i.e. the operator f of the model, has the form

$$f: \frac{d^2y}{dt^2} + c_1 \frac{dy}{dt} = c_2 u(t),$$

where y is the covered distance while $y(0) = \dot{y}(0) = 0$, i.e. the start is made at a zero time moment with a zero speed, c_1 and c_2 are the parameters of braking and speeding up to be determined at the stage of identification (or experiment design); $u(t)$ is the scalar control ($q = 1$) of acceleration realized with the aid of the throttle (gas pedal) and the brakes. It is the force applied to the car $a_1 \leqslant u(t) \leqslant a_2$, where $a_1 < 0$ is the maximal force of braking (it depends on the quality of brakes and the condition of the road); $a_2 > 0$ is the maximal force of speeding up (it depends on the power of the engine, the quality of gasoline, etc.).

Suppose T is the time required to pass from one cross-road to another. This is a functional whose value is determined by the selected function of control $u(t)$. To find the value of the functional, it is necessary to solve the equation of the model with the boundary conditions

$$y(0) = \dot{y}(0) = \dot{y}(T) = 0; \quad y(T) = L,$$

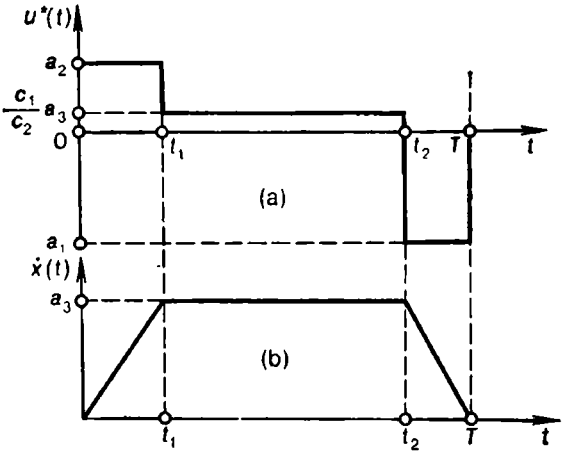

Fig. 6.1.1. An example of optimal control

where L is the distance between the cross-roads. The obtained value of the functional $T = T(u(t))$ is the extremal (minimal) objective.

Now problem (6.1.6) for this case may be presented as

$$T(u(t)) \to \min_{u(t)\in\Omega},$$

$$\Omega: \begin{cases} a_1 \leqslant u(t) \leqslant a_2, \\ \ddot{y} + c_1\dot{y} = c_2 u(t), \\ y(0) = \dot{y}(0) = \dot{y}(T) = 0, \ y(T) = L, \\ \dot{y}(t) \leqslant a_3. \end{cases}$$

We added here to the stated constraints the speed limit ($a_3 = 60$ km/h) which is established by the traffic regulations. The optimal control is the function $u^*(t)$. It is difficult to solve this problem, but the common sense prompts the optimal solution which is used by taxi-drivers. It goes as follows: first 'put the foot all the way down' on the trottle, i.e. $u(t) = a_2$ until $\dot{y} = a_3$, then keep the speed at the maximum level a_3 allowed by the rules (the equation of the model gives $u = (c_1/c_2)\, a_3$ for this period), then brake hard ($u = a_1$) until a complete stop, i.e. make $\dot{y} = 0$.

The diagram of such an optimal control $u^*(t)$ and the corresponding speed diagram are shown in Fig. 6.1.1*a*, *b*.

Clearly, the problem to synthesize the optimal control $u^*(t)$ is reduced to the determination of only two time moments t_1 (the end of acceleration) and t_2 (the start of braking). Let us note that the medium X (the quality of the road pavement in this case)

may change the braking length, i.e. $a_1 = a_1(X)$, and the maximal allowed speed under the conditions X, i.e. $a_3 = a_3(X)$.

The obtained solution is optimal for the taxi-driver, but obviously unacceptable for the passenger who will be forced to endure great overload (especially when braking). Therefore, in the interests of the passenger, it is necessary to lower the overload to an acceptable level $a_1^* > a_1$ (let us recall that these are negative values). The value of a_1^* is determined by the comfortable conditions for the passenger.

The *control of the dynamic object* is reduced, therefore, to the solution of variational problem (6.1.6) where the sought control is a function of time $\mathbf{U} = \mathbf{U}(t)$, while Q, G, and H, generally speaking, are some given functionals of control $\mathbf{U}(t)$.

Obviously, this problem is essentially more involved than the problem of mathematical programming. Therefore, the solution of the variational problem with arbitrary functionals is usually reduced to the solution of problems of mathematical programming. It is sufficient for that to make the control function $\mathbf{U}(t)$ parametric.

It can be done in two ways:

(1) by means of expansion $\mathbf{U}(t)$ into a series in terms of a specified system of functions $\{\Phi_i(t)\}$:

$$\mathbf{U}(t) = \sum_{i=1}^{k} c_i \Phi_i(t),$$

which results in the problem of finding the parameters $c_1, \ldots, c_k$; this approach is known as parametrical;

(2) by presenting the function $\mathbf{U}(t)$ by its samples $\mathbf{C}_1, \ldots, \mathbf{C}_k$ in some specified time moments, where $\mathbf{C}_i = \mathbf{U}(t_i)$ $(i = 1, \ldots, k)$ are new sought parameters of control; this approach is commonly called non-parametrical.

In both cases, substituting these expressions into (6.1.7), we get after some obvious transformations a problem of mathematical programming of the form

$$\begin{aligned} &\mathbf{Q}(\mathbf{X}, \mathbf{U}) \to \min_{\mathbf{U} \in \Omega}, \\ &\Omega: \begin{cases} \mathbf{G}(\mathbf{X}, \mathbf{U}) = 0, \\ \mathbf{H}(\mathbf{X}, \mathbf{U}) \geqslant 0, \end{cases} \end{aligned} \tag{6.1.8}$$

where $\mathbf{U} = (C_1, \ldots, C_k)$ is the vector of the sought parameters, while $\mathbf{Q}$, $\mathbf{G}$, and $\mathbf{H}$ are some specified vector functions of these parameters.

Consequently, the solution of any problem of the control synthesis may be reduced, in the long run, to the solution of a problem of mathematical programming. Let us consider various types of the problems of this kind.

6.2. Classification of Problems of Mathematical Programming

The form and the number of functions $\mathbf{Q}$, $\mathbf{G}$, and $\mathbf{H}$ lead to different types of problems of mathematical programming, and different methods for their solution (that is why the classification is presented). Depending on the number k_3 of the minimized functions q_l ($l = 1, \ldots, k_3$), *monocriterial* ($k_3 = 1$) and *multicriterial* ($k_3 > 1$) problems are distinguished. In the latter case they are sometimes called the problems of vector optimization (the problems of optimization with a vector criterion). A specific feature of multicriterial problems of mathematical programming is the necessity to introduce additional data on the object and the criteria without which the problem has no solution. The methods to solve the multicriterial problems are discussed in the 'Supplement'.

Here we shall consider only the monocriterial problems ($k_3 = 1$) for which $\mathbf{Q}$ is a scalar function, i.e. $\mathbf{Q} = Q$.

Depending on the form and the character of functions Q, $\mathbf{G}$, and $\mathbf{H}$ the following problems of mathematical programming are distinguished.

1. In case Q, $\mathbf{G}$, and $\mathbf{H}$ are linear with regard to $\mathbf{U}$, we have the problem of *linear* programming:

$$Q(\mathbf{U}) = \sum_{l=1}^{q} \alpha_l u_l \to \min_{\mathbf{U} \in \Omega}, \qquad (6.2.1)$$

where

$$\Omega: \begin{cases} \sum\limits_{k=1}^{q} a_{ik} u_k = c_i, & i = 1, \ldots, k_1 < q, \\ \sum\limits_{k=1}^{q} b_{jk} u_k \geqslant d_j, & j = 1, \ldots, k_2. \end{cases}$$

The parameters a_l, a_{ik}, b_{jk}, d_j, and c_i are determined by the factors of the medium X and the object F.

The process to solve the problem of linear programming is naturally underlied by the linearity of expression (6.2.1), which is realized by the so-called simplex method.

The optimal cocktail problem discussed above (see § I.5) was a problem of linear programming. Another example of such problems is the problem of optimal mixture. Let us consider it.

Example. There are q kinds of basic raw materials A_1, A_2, ..., A_q, each consisting of n components

$$A_k = (a_{1k}, a_{2k}, \ldots, a_{nk}), \quad k = 1, \ldots, q$$

(here a_{ik} is the relative content of the i-th component in the k-th raw material). It is required to make a mixture of a specified composition $c_1, \ldots, c_n$, where c_i is the relative content of the i-th component in the mixture.

The mixture is characterized by the vector of the composition $U = (u_1, \ldots, u_q)$, where u_k is the relative amount of the k-th raw material $\left(\sum_{k=1}^{q} u_k = 1\right)$ The cost of the raw materials is determined by the prices $a_1, \ldots, a_q$, where a_l is the price of the l-th material unit ($l = 1, \ldots, q$). The cost of a unit mixture should be minimal.

Therefore, the problem may be reduced to minimization of the linear function of the cost of the mixture

$$Q(U) = \sum_{l=1}^{q} a_l u_l \to \min_{U \in \Omega},$$

where the set Ω of the admissible values of the composition U is determined by the requirements to the mixture composition:

$$\Omega: \begin{cases} \sum_{k=1}^{q} a_{ik} u_k = c_i, & i = 1, \ldots, n, \\ u_k \geqslant 0, & k = 1, \ldots, q. \end{cases}$$

The inequality $u_k \geqslant 0$ expresses an obvious requirement: the amount of raw materials in the mixture cannot be negative. The constraints Ω include n equalities (G) and q inequalities (H); $n < q$ (in case $n = q$ the problem degenerates into a non-extremal one). This is a typical problem of linear programming.

The methods to solve the problems of linear programming are considered in § 6.4.

2. Another kind of problems of mathematical programming is the *problems of convex programming*; they are of a

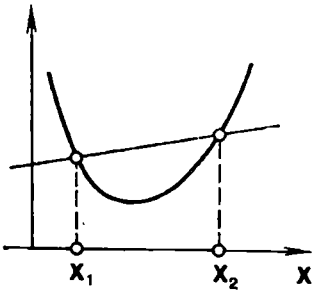

Fig. 6.2.1. To the definition of a convex function

more general nature and may be formulated as follows.

It is necessary to minimize a downwards convex function

$$Q(\mathbf{U}) \to \min_{\mathbf{U} \in \Omega}, \tag{6.2.2}$$

where the area Ω of admissible states is specified by a system of inequalities

$$\Omega : h_i(\mathbf{U}) \geqslant 0, \quad i = 1, \ldots, k_2. \tag{6.2.3}$$

Here $h_i(\mathbf{U})$ are downwards convex functions, which guarantees convexity of the area Ω.

Let us note that a function $f(\mathbf{X})$ is called downwards convex if any segment connecting any two points of the function lies higher than its values (Fig. 6.2.1). This definition may be written formally in the following way. If for any point $\mathbf{X}$ of the segment between the points $\mathbf{X}_1$ and $\mathbf{X}_2$, i.e. for any

$$\mathbf{X} = \lambda \mathbf{X}_1 + (1 - \lambda)\mathbf{X}_2, \quad 0 \leqslant \lambda \leqslant 1 \tag{6.2.4}$$

the inequality

$$f(\mathbf{X}) \leqslant \lambda f(\mathbf{X}_1) + (1 - \lambda) f(\mathbf{X}_2) \tag{6.2.5}$$

holds true, then the function is downwards convex. It is easy to see that the right-hand side of inequality (6.2.5) is the linear approximation of the function $f(\mathbf{X})$ in terms of its two values in the points $\mathbf{X}_1$ and $\mathbf{X}_2$ (see the straight line in Fig. 6.2.1).

It is assumed that a convex area Ω is a locus of the following kind. If one takes two points of the locus, then the segment of the straight line connecting these two

points belongs to the locus. Formally: if $\mathbf{X}_1, \mathbf{X}_2 \in \Omega$, then $\mathbf{X} \in \Omega$, where $\mathbf{X}$ is defined (6.2.4) as a point between $\mathbf{X}_1$ and $\mathbf{X}_2$.

The convex functions are a rather wide class of functions and they may comparatively well describe problems (6.1.7) of the control synthesis.

An essential feature of the problems of convex programming is the locality, i.e. the presence of only one extremum. This means that solving the problems of this kind one may be satisfied with finding a local extremum. The properties of convexity of the minimized function Q and the area Ω are used in solving the problems of convex programming. Let us note that the problems of linear programming are a particular case of convex programming. This is easy to prove since any linear function meets the requirements of convexity. A method to solve the problems of convex programming is described in § 6.4.

3. The convexity requirements are far from always met in the problems of control synthesis. In the general case, when Q, G, and H are arbitrary non-linear functions, the problem is that of *non-linear programming.*

This is a rather general form of mathematical programming. Generally speaking, it is not possible to guarantee here a coincidence of the local and global extrema, which frequently causes serious difficulty in solving the problems of this kind belonging, if that is the case, to the multiextremal class. However, even if the problem is monoextremal, there are no special methods for solving the problems of non-linear programming. All methods applied for that are the optimum seeking ones (some of them are considered in the 'Supplement').

The typical problems of non-linear programming are the problems of optimal design. When a new circuit, structure or system is designed, there is always a definite extremal objective, for instance, efficiency, reliability, operational speed, the mass of the structure, etc. Optimal design is a process of finding such parameters $u_1, \ldots, u_q$ of the structure that make the objective extremal, and not arbitrarily, but observing certain constraints. The type of constraints results from some strict requirements. For instance, there is a need to use certain structural elements. This is a constraint of the type of equalities. Some other

constraints may be inequalities. For instance, temperature and arising tensions should not exceed specified standard limits, etc. Consequently, the problems of the optimal design are typical problems of non-linear programming, because both the constraints and the quality function are commonly of clearly pronounced non-linear nature.

4. Very commonly the control **U** is discrete, i.e. the set of admissible controls Ω is finite and discrete. For instance, if the components of the vector of control u_i ($i = 1, \ldots, q$) may take only integers, the problem is that of *integer programming.*

Example. A problem of integer programming is the problem to control a system of some interacting units when each i-th unit may either operate ($u_i = 1$) or not operate ($u_i = 0$). The efficiency of the system should be maximized at a specified energy resource R:

$$Q(\mathbf{U}) \to \max_{\mathbf{U} \in \Omega},$$

$$\Omega: \begin{cases} \sum_{i=1}^{q} a_i u_i \leqslant R, \\ u_i = \{0;\ 1\}, \quad i = 1, \ldots, q, \end{cases}$$

where a_i is the cost of the operation of the i-th unit.

In the simplest case, if the units do not interact,

$$Q(\mathbf{U}) = \sum_{i=1}^{q} b_i u_i,$$

where b_i is the efficiency of the i-th unit.

That is an integer problem of linear programming.

The solution of the problems of this kind is sometimes very difficult. They are going to be considered in § 6.4.

5. If Q, **G**, and **H** are stochastic characteristics, then the problem is that of *stochastic programming.*

Let us recall that the most common stochastic characteristics are:

the probability

$$P(\mathbf{X} \in S) = \int_S p(\mathbf{X})\, d\mathbf{X}, \tag{6.2.6}$$

the mathematical expectation

$$M(\mathbf{X}) = \int \mathbf{X} p(\mathbf{X})\, d\mathbf{X}, \tag{6.2.7}$$

and the dispersion

$$D(\mathbf{X}) = \int [\mathbf{X} - M(\mathbf{X})]^2 p(\mathbf{X})\, d\mathbf{X} \qquad (6.2.8)$$

etc., where $p(\mathbf{X})$ is the density of the distribution of the random factor $\mathbf{X}$, and the integral is multiple: its multiplicity is equal to the dimension of $\mathbf{X}$.

An example of the problem of stochastic programming is the solution of basic problem (6.1.8) with regard to the mean values over the random state of the medium $\mathbf{X}$, i.e. the solution of the problem 'on the average'. Then (6.1.8) is to be transformed into

$$\overline{Q}(\mathbf{U}) \to \min_{\mathbf{U} \in \Omega},$$

$$\Omega : \begin{cases} \overline{\mathbf{G}}(\mathbf{U}) = 0, \\ \overline{\mathbf{H}}(\mathbf{U}) \geqslant 0, \end{cases}$$

where

$$\overline{Q}(\mathbf{U}) = \int Q(\mathbf{X}, \mathbf{U})\, p(\mathbf{X})\, d\mathbf{X}; \quad \mathbf{G}(\mathbf{U}) = \int \mathbf{G}(\mathbf{X}, \mathbf{U})\, p(\mathbf{X})\, d\mathbf{X}$$

$$\overline{\mathbf{H}}(\mathbf{U}) = \int \mathbf{H}(\mathbf{X}, \mathbf{U})\, p(\mathbf{X})\, d\mathbf{X},$$

$p(\mathbf{X})$ is a known density of distribution of the state $\mathbf{X}$ of the medium. This is a function of n variables $x_1, \ldots, x_n$ (the inputs of the object) with the natural constraints

$$p(\mathbf{X}) \geqslant 0, \quad \int p(\mathbf{X})\, d\mathbf{X} = 1.$$

The integrals both here and above are n-fold and taken over the whole space of the inputs $\{\mathbf{X}\}$.

The solution of the problem of stochastic programming is reduced to finding some stochastic characteristics, i.e. to the computation of multiple integrals of the form

$$J = \int f(\mathbf{X}, \mathbf{U})\, p(\mathbf{X})\, d\mathbf{X}, \qquad (6.2.9)$$

which is a difficult calculation problem itself. However, this difficulty may be negotiated, if we use instead of J its estimate according to the Monte Carlo method:

$$\hat{J}_N = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{X}_i, \mathbf{U}), \qquad (6.2.10)$$

where $\mathbf{X}_i$ $(i = 1, \ldots, N)$ are independent random vectors generated in compliance with the specified density of the distribution p (X). Indeed, J is the mean of the function f:

$$J = M_{\mathbf{X}} f(\mathbf{X}, \mathbf{U}),$$

where $M_{\mathbf{X}}$ is the mathematical expectation over X. But expression (6.2.10) is an estimate of that mean over the basis of N, and therefore it can serve as an approximate value of (6.2.9).

Estimate (6.2.10) obtained with the aid of the Monte Carlo method is a random value characterized by the following stochastic properties. Its mean coincides with J: $M(\hat{J}_N) = J$, i.e. the estimate $\hat{J}_N$ is unbiased, the limit $\lim_{N\to\infty} \hat{J}_N = J_N$, i.e. the estimate is consistent, and its dispersion $D(\hat{J}_N) = c/N$ where c is a constant. This implies that the estimate by the Monte Carlo method is 'good', i.e. it coincides with the sought one on the average at any large enough N.

These properties make it possible to write an expression relating the estimate to the exact value in the form

$$\hat{J}_N = J + \varepsilon(c/N) \tag{6.2.11}$$

where $\varepsilon(c/N)$ is a random value with a zero mean and the dispersion c/N.

Expression (6.2.11) means that the Monte Carlo estimate differs from the exact one by a random value equal to zero on the average, and whose dispersion decreases with the increase of N as its reciprocal $\frac{1}{N}$. This allows for interpreting $\hat{J}_N$ as the result of calculating J with a random deviation related to the method of calculation. Therefore, the problem of stochastic programming may be written in the standard form (6.1.8) with the reservation that in the process of solution each of the functions has an imposed random deviation whose properties, generally speaking, may be unknown. The solution of this problem presents some natural difficulties that may be overcome by using the seeking methods of the stochastic approximation type (see the 'Supplement').

6.3. Analysis of Problems of Mathematical Programming

The main method to solve the problems of mathematical programming is the so-called *recurrent method.* It consists in finding consequtive approximate solutions to the problem

$$\mathbf{U}_0,\ \mathbf{U}_1,\ \ldots,\ \mathbf{U}_N,\ \ldots, \tag{6.3.1}$$

which have the following important properties:

(1) each next member of the sequence is in a certain sense (from the point of view of solving the problem) more preferable than the preceding member, i.e. $\mathbf{U}_{N+1} \succ \mathbf{U}_N$, where $\succ$ is the preference symbol;

(2) at large N the sequence tends to a value $\mathbf{U}^*$ that can be considered to be the solution of the problem. It should be noted that commonly the problems of mathematical programming do not require any strict solution since the initial data for them are always approximate. Therefore, we can only speak of an approximation to a rational solution. It is only in a particular case that $\mathbf{U}^*$ may be the exact solution of the problem.

Let us consider some methods of formal confirmation of the said preference. There are various methods to do that. Here are some examples helping introduce obvious preferences.

1. $\mathbf{U}_{N+1} \succ \mathbf{U}_N$ if $\mathbf{U}_N, \mathbf{U}_{N+1} \in \Omega$ and $Q(\mathbf{U}_{N+1}) < Q(\mathbf{U}_N)$, i.e. out of the admissible solutions (those satisfying Ω) the one, whose quality function $Q(\mathbf{U})$ is less, is better.

2. If $Q(\mathbf{U}_{N+1}) = Q(\mathbf{U}_N)$ and $\mathbf{U}_{N+1},\ \mathbf{U}_N \in \Omega$, then $\mathbf{U}_{N+1} \sim \mathbf{U}_N$, i.e. the solutions are equivalent ($\sim$ is the equivalence sign).

3. $\mathbf{U}_{N+1} \succ \mathbf{U}_N$ if $\mathbf{U}_{N+1} \in \Omega$ while $\mathbf{U}_N \notin \Omega$, i.e. an admissible solution is always more preferable.

It is possible to add more to the list. However, the determination of preferences is not the final solution to the problem of synthesizing an algorithm to solve the initial problem for any situation arising in the process of control. It is easy to note that the described preferences include far from all the situations that may appear in the

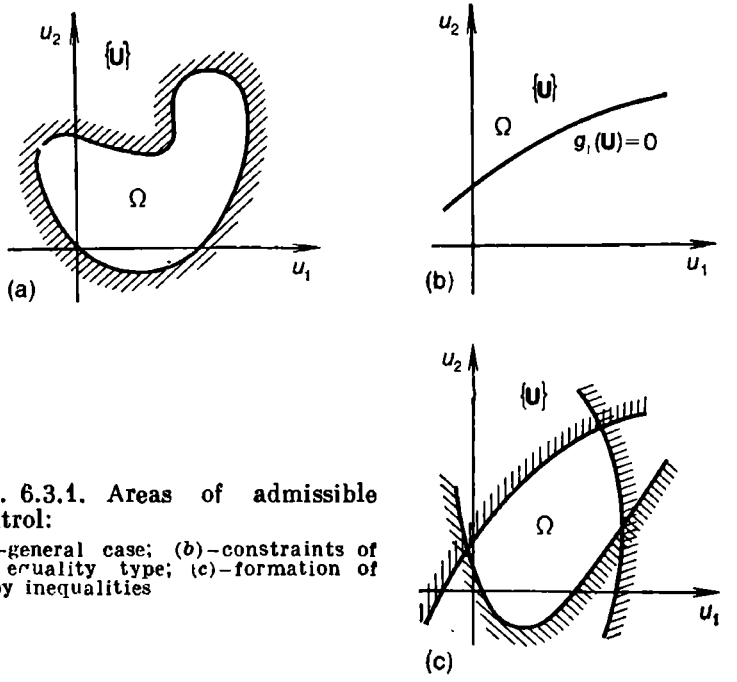

Fig. 6.3.1. Areas of admissible control:
(*a*)–general case; (*b*)–constraints of the equality type; (c)–formation of Ω by inequalities

control of an object. The problem is solved by means of an algorithm. Let us consider the algorithm to solve the problem of mathematical programming to be the operator φ linking the subsequent solutions:

$$\mathbf{U}_N = \mathbf{U}_{N-1} + \varphi(I_N, \mathbf{U}_{N-1}, Q_{N-1}, \mathbf{G}_{N-1}, \mathbf{H}_{N-1}).$$

Here φ is the algorithm indicating how the control should change at the existing values of $\mathbf{U}_{N-1}$, $Q_{N-1} = Q(\mathbf{U}_{N-1})$, $\mathbf{G}_{N-1} = \mathbf{G}(\mathbf{U}_{N-1})$, $\mathbf{H}_{N-1} = \mathbf{H}(\mathbf{U}_{N-1})$, and the acquired additional information I_N (all that is the situation reached in the process of control).

Let us consider the space of the solutions {U}. The conditions Ω limit a certain area of admissible solutions $\mathbf{U} \in \Omega$ (Fig. 6.3.1*a*) within this space. The dimension of the area Ω is $(q - k_1)$ since each of the constraints of the type of the equalities $g_i(\mathbf{U}) = 0$ decreases the dimension of the problem by unity (Fig. 6.3.1*b*). The constraints of the type of inequalities $h_j(\mathbf{U}) \geqslant 0$ delineate the borders of the area of the admissible solutions Ω (Fig. 6.3.1*c*).

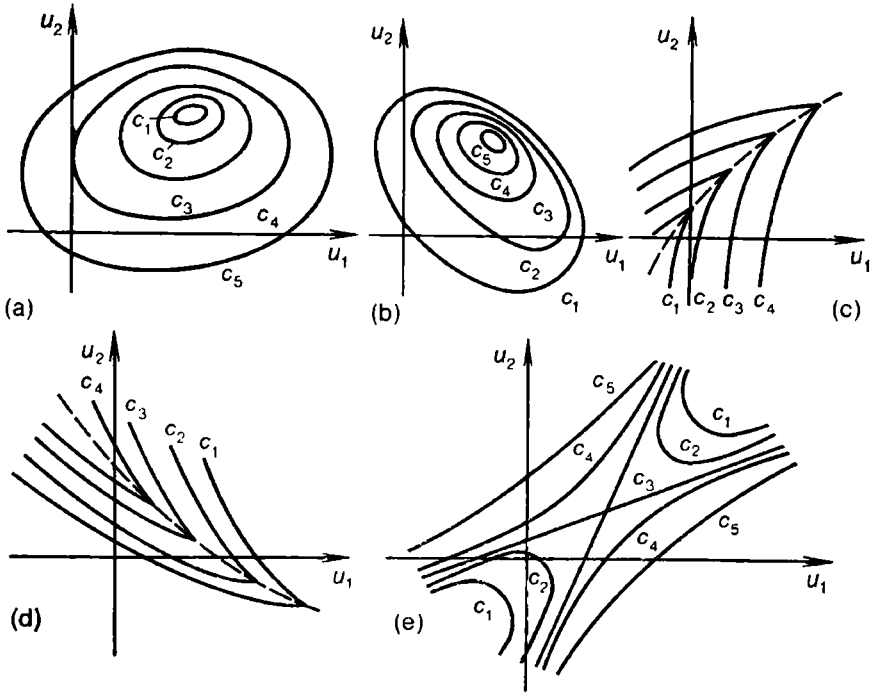

Fig. 6.3.2. Examples of the 'reliefs' of the quality functions ($c_1 < c_2 < c_3 \ldots$)

Clearly, the problem under discussion is unambiguously determined by the area Ω and the scalar quality function Q (U). The function forms in Ω a scalar field which in the case of $q = 2$ is convenient to be characterizd by the lines of equal level for the minimized function. The equations for these lines

$$Q\,(\mathbf{U}) = c = \text{const} \qquad (6.3.2)$$

are obtained at various values of the parameter c. It is easy to note that the lines of equal level on the plane of the controlled parameters $\{u_1, u_2\}$ are similar to the topographical equal level lines (e.g. above sea level) on a geographical map. That is why topographical terms are used in describing optimum problems. Here are some topographical terms widely applied in the optimization theory and practice: 'equal level lines' (6.3.2) forming the 'relief' of the function Q (U), the relief may be diverse:

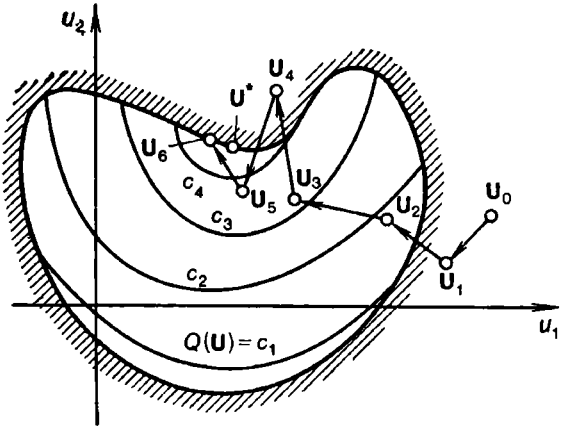

Fig. 6.3.3. Example of a trajectory in solving a problem of mathematical programming ($c_1 > c_2 > c_3 \ldots$)

a 'hole', a 'pit', a 'depression' (Fig. 6.3.2*a*), a 'hill' or 'peak' (Fig. 6.3.2*b*), a 'ravine' or 'cleft' (Fig. 6.3.2*c*), a 'ridge' (Fig. 6.3.2*d*), a 'pass' or 'saddle' (Fig. 6.3.2*e*), a 'plateau', etc. The analogy covers the solution process for the optimum problems too. Thus, a movement in the direction of decreasing in the quality function is called 'descent', a movement in the direction of increasing it is called 'ascent'.

Naturally, such demonstrative analogy may also be used for the multidimensional optimum problems at $q > 2$. In this case the equal level lines become at $q = 3$ surfaces and at $q > 3$ hypersurfaces of equal level. Now the recurrent process of solving a problem of mathematical programming may be presented as moving in the q-dimensional space $\{\mathbf{U}\}$ from point to point in sequence (6.3.1) (Fig. 6.3.3).

The problems of mathematical programming may be solved in various ways. It is possible to use a priori data on the structure of the functions Q, $\mathbf{H}$, and $\mathbf{G}$, i.e. to develop algorithms taking into account the specific features of these functions by means of special methods (some of them are going to be considered in § 6.4). But if the object of control is complex, it is difficult to find out those specific features and take them into account, and it is much easier to do otherwise: to organize a process of seeking

the solution U* using only the values of the functions, not the data on their structure. Such an approach is called the optimum seeking method. Its nature is universal. Some of the most widely applied methods of optimum seeking are described in the 'Supplement'.

Let us consider in what manner the a priori data on the problem may be 'put into' the algorithm for its solution.

6.4. Some Special Methods to Solve Problems of Mathematical Programming

Let us consider some of the methods used to solve the problems of mathematical programming. The specific features of those methods are first of all determined by the specific structure of the functions Q, **G**, and **H**.

The simplex method. This is a method to solve the problems of linear programming when the functions Q, **G**, and **H** are linear with regard to **U**:

$$Q(\mathbf{U}) = \sum_{l=1}^{q} a_l u_l \to \min_{\mathbf{U} \in \Omega}, \tag{6.4.1}$$

where

$$\Omega : \sum_{k=1}^{q} b_{jk} u_k \geqslant c_j, \quad j = 1, \ldots, k_2.$$

This is different from (6.2.1) where also constraints of the type of equalities are present. But the linearity allows for their easy elimination thus decreasing the number of independent variables. On the other hand, there is a formal method to get rid of these constraints by means of substituting $q_i(\mathbf{U}) = 0$ with two constraints: $q_i(\mathbf{U}) \geqslant 0$ and $-q_i(\mathbf{U}) \geqslant 0$. Let us assume that one of these operations has been carried out and the problem of linear programming acquired the form (6.4.1).

The number of constraints forming the area of admissible solutions Ω in the problems of linear programming should always exceed the number of variables, i.e.

$$k_2 > q, \tag{6.4.2}$$

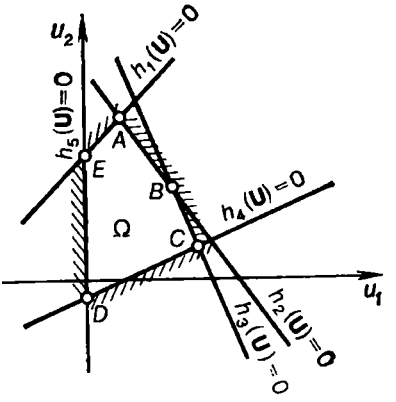

Fig. 6.4.1. Example of forming a convex polyhedron Ω in a problem of linear programming

otherwise the problem, generally speaking, may not have a solution. Let us note that usually q of the constraints have the form

$$u_i \geqslant 0, \quad i = 1, \ldots, q, \tag{6.4.3}$$

i.e. the solution is sought in the positive quadrant (it is related to the physical meaning of the variables: u_i is commonly an amount of something). Let us introduce a designation

$$h_j = h_j(\mathbf{U}) = \sum_{k=1}^{q} b_{jk} u_k - c_j, \quad j = 1, \ldots, k_2. \tag{6.4.4}$$

Then the equation of the area Ω may be written in the standard form

$$\Omega : h_j = h_j(\mathbf{U}) \geqslant 0, \quad j = 1, \ldots, k_2. \tag{6.4.5}$$

It is easy to see that the area is a convex polyhedron. Indeed, each of the conditions $h_j(\mathbf{U}) \geqslant 0$ defines a half-space. The intersection of all these half-spaces may only produce either a convex or empty (if the constraints are contradicting) polyhedron. An example of the formation of such a convex polyhedron for $q = 2$ is shown in Fig. 6.4.1.

Owing to linearity of $Q(\mathbf{U})$, the solution of the problem of linear programming (6.4.1) lies on the surface of that polyhedron. Indeed, it cannot be within it since other-

wise the constraints for Ω could be removed. But then it would be $U^* \to \infty$ because of the linearity of $Q(U)$; therefore

$$U^* \in \bar{\Omega}, \tag{6.4.6}$$

where $\bar{\Omega}$ is the surface of the polyhedron Ω. Further, it is easy to see that the solution is always in one of the vertices or on one of the faces. In the latter case any point of this face, its vertices in that number, may be assumed to be the solution of problem (6.4.1) since $Q(U)$ is linear. Therefore, to solve the problem, it is sufficient to consider the vertices of the polyhedron Ω.

And the last point: since each vertex of the polyhedron Ω is determined, in the general case, by q out of k_2 constraints (6.4.5), then the q constraints in each vertex are equal to zero, the rest of them being positive (to be more precise, non-negative). This allows for coding each of the vertices by a vector

$$H = (h_1, \ldots, h_{k_2}), \tag{6.4.7}$$

whose components are equal to the values of the constraints (6.4.5) for this vertex. Vector (6.4.7) at that has at least q zero components, the rest of them being non-negative.

For instance, in the case shown in Fig. 6.4.1 there are the following codes for the vertices $A, \ldots, E$:

$$H_A = (0, 0, h_3, h_4, h_5),$$
$$H_B = (h_1, 0, 0, h_4, h_5),$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$H_E = (0, h_2, h_3, h_4, 0).$$

The simplex method to solve the problems of linear programming is to pass over the faces of the polyhedron Ω to the neighbouring (contiguous) vertices where the value of the quality function Q is less than in the preceding vertex. The algorithm of the simplex method includes the following operations at each step.

1. Determine all the vertices $U_N^{(i)}$ $(i = 1, \ldots, k_N)$ contiguous to the vertex U_N at the N-th step (k_N is the number of the vertices contiguous to U_N).

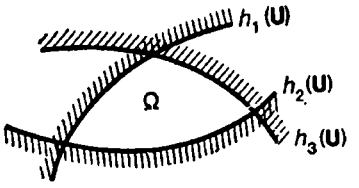

Fig. 6.4.2. Example of an area formed by downwards convex constraints

2. Select the one where the quality function is minimal:

$$Q(U)_N^{(j)} = \min_{i=1, \ldots, k_N} Q(U_N^{(i)}).$$

3. Pass over to that vertex, i.e.

$$U_{N+1} = U_N^{(j)},$$

then start the next $(N + 1)$-th step, i.e. determine the vertices contiguous to U_{N+1}, and so on (see step 1).

The operation of the algorithm comes to an end in case among all the vertices contiguous to U_N there is no vertex with less a value of the quality function than that for the preceding one U_N. In this case $U^* = U_N$.

The main difficulty of the simplex method is the determination of the contiguous vertices. The difficulty may be negotiated by means of the so-called simplex tables mentioned in any textbook on linear programming (see 'Comments' to this chapter).

The method of possible directions. This method is applied to solve the problems of convex programming of the form

$$Q(U) \to \min_{U \in \Omega}, \quad \ldots\ldots\ldots\ldots \quad (6.4.8)$$

where

$$\Omega : h_j(U) \geqslant 0, \; j = 1, \ldots, k_2;$$

$Q(U)$ and $h_j(U)$ are downwards convex functions with property (6.2.5). The area Ω at that is convex (Fig. 6.4.2) because the intersection of convex areas produces a convex area.

The solution to problem (6.4.8), generally speaking, may be both within the area Ω and on its margin.

Let us introduce the concept of a possible direction. Suppose U_N is a certain point belonging to Ω ($U_N \in \Omega$).

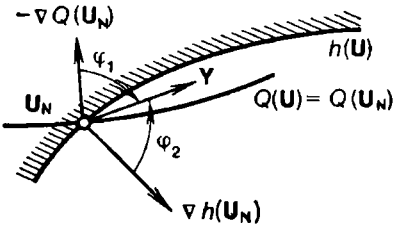

Fig. 6.4.3. Example of an area for which the antigradient direction is impossible

Then a possible direction is determined by such a vector $\mathbf{Y}$ along which the quality function $Q(\mathbf{U})$ decreases without violating the constraints Ω, i.e.

$$Q(\mathbf{U}_N + \alpha\mathbf{Y}) < Q(\mathbf{U}_N), \tag{6.4.9}$$

where $(\mathbf{U}_N + \alpha\mathbf{Y}) \in \Omega$ and $\alpha > 0$ is small.

The solution of the problem of finding a possible direction $\mathbf{Y}$ presents the main difficulty while realizing this method. If the preceding point $\mathbf{U}_N$ is inside the area Ω (i.e. not on its margin), the best is the antigradient direction

$$\mathbf{Y} = -\nabla Q(\mathbf{U}_N).$$

But if $\mathbf{U}_N$ belongs to the margin of the area Ω, then the antigradient direction $-\nabla Q(\mathbf{U}_N)$ may prove to be impossible (Fig. 6.4.3, where ∇h is the vector of the gradient for the constraints $h(\mathbf{U}) \geqslant 0$ in the point $\mathbf{U}_N$). As it is well seen, any possible direction $\mathbf{Y}$ requires that its angles φ_1 and φ_2 with the vectors $-\nabla Q(\mathbf{U}_N)$ and $\nabla h(\mathbf{U}_N)$ be less than $\pi/2$, i.e. $\varphi_i < \pi/2$ $(i = 1, 2)$. This requirement can always be met if the vectors $-\nabla Q$ and ∇h are not exactly opposite.

We can demand more. Since there is a whole set of possible directions $\mathbf{Y}$, it is natural to select one of them which could be optimal in some sense. Such a direction $\mathbf{Y}^*$ is the one along which the quality function $Q(\mathbf{U})$ decreases most intensely, i.e. the angle φ_1 is minimal while the departure from the constraint $h(\mathbf{U}) = 0$ is the greatest, i.e. the angle φ_2 is also minimal. These contradicting requirements (since the sum $\varphi_1 + \varphi_2$, as seen in Fig. 6.4.3, is constant) can be satisfied by solving the pertinent problem of linear programming (see the books concerning convex programming mentioned in the 'Comments' to this chapter).

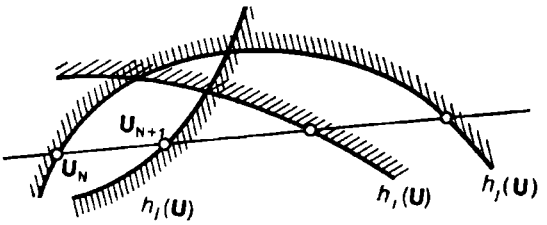

Fig. 6.4.4. To finding the coefficient α'_l

Having found the optimal possible direction $\mathbf{Y}^*$, we should organize a descent along it, i.e. find a value of α^* in $\mathbf{U} = \mathbf{U}_N + \alpha^*\mathbf{Y}$ more than which α cannot be since either the quality function $Q(\mathbf{U})$ is going to increase or one of the constraints $h_j(\mathbf{U}) \geqslant 0$ $(j = 1, \ldots, k_2)$ is violated. Therefore we should solve the problems

$$Q(\mathbf{U}_N + \alpha\mathbf{Y}^*) \to \min_{\alpha>0} \Rightarrow \alpha'_0 > 0, \tag{6.4.10}$$

$$h_j(\mathbf{U}_N + \alpha\mathbf{Y}^*) = 0 \Rightarrow \alpha'_j > 0, \quad j = 1, \ldots, k_2$$

and choose the least α'_j $(j = 0, \ldots, k_2)$:

$$\alpha^* = \min_{j=0, \ldots, k_2} \alpha'_j = \alpha'_l. \tag{6.4.11}$$

The need for such a selection is well seen in the example in Fig. 6.4.4, where $\alpha'_0 > \alpha'_l$.

This is the end of the N-th stage in solving the problem of convex programming $\mathbf{U}_{N+1} = \mathbf{U}_N + \alpha^*\mathbf{Y}^*$, after which the next $(N + 1)$-th stage should come where the gradient of the l-th constraint (6.4.11) is to be calculated to find another possible direction.

The methods to solve problems of integer programming. Let us consider the methods to solve problems of integer programming

$$Q(\mathbf{U}) \to \min_{\mathbf{U}\in\Omega_d} \Rightarrow \mathbf{U}^*_d, \tag{6.4.12}$$

$$\Omega_d : \begin{cases} h_j(\mathbf{U}) \geqslant 0, & j = 1, \ldots, k_2, \\ u_i \in \{u^{(1)}{}_i, \ldots, u^{(l_i)}{}_i\}, & i = 1, \ldots, q, \end{cases}$$

where l_i is the finite number of discrete values for the i-th component of the control vector $\mathbf{U} = (u_1, \ldots, u_q)$. It follows that the set Ω_d of the admissible values for

the control $\mathbf{U}$ is finite and has $\prod_{i=1}^{q} l_i$ elements, which makes an exhaustive search at large l_i and q impossible, i.e. rules out 'the frontal attack' on problem (6.4.12). Therefore let us consider other techniques.

Suppose that for certainty $u_i \in \{1, \ldots, l_i\}$. The simplest method is to solve the continuous problem

$$Q(\mathbf{U}) \to \min_{\mathbf{U} \in \Omega} \Rightarrow \mathbf{U}^*,$$

$$\Omega : h_j(\mathbf{U}) \geqslant 0, \quad i = 1, \ldots, k_2,$$

which is obtained from (6.4.12) by way of eliminating the constraint on the components of the vector $\mathbf{U}$ to be integers, and then selecting in the nearest vicinity of the solution $\mathbf{U}^*$ an integer vector $\mathbf{U}'_d$ meeting the constraints Ω at the minimal value of the quality function Q. This requires first to find the nearest integer point by means of rounding off the coordinates to integers (e.g. $2.1 \to 2$, $3.8 \to 4$), and then to check for the constraints. If the constraints are met, this point is called the solution to the problem of integer programming, i.e. $\mathbf{U}'_d = \mathbf{U}^*_d$ (an example of such a point is shown in Fig. 6.4.5*a*). Now in case the nearest integer point $\mathbf{U}'_d$ is beyond the admissible area Ω, i.e. $\mathbf{U}'_d \notin \Omega$ (this is demonstrated in Fig. 6.4.5*b*), it is necessary then to consider its integer vicinity (it is delineated by the dashes in Fig. 6.4.5*b*) containing $(3^q - 1)$ points. Out of these points we should single out first a subset of those meeting the constraints Ω. Suppose it includes the points $\mathbf{U}_1, \ldots, \mathbf{U}_k$. The basis for the selection of one of them $\mathbf{U}^*_d$ is the minimal value of the quality function, i.e.

$$Q(\mathbf{U}^*_d) = \min_{i=1, \ldots, k} Q(\mathbf{U}_i).$$

However, this method far from always brings the correct solution. This case is shown in Fig. 6.4.5*c* where the solution of the integer problem is in the point **A**, and not in **B** obtained by the given method (here the parallel lines are the equal level lines). This circumstance brings about the use of other methods to solve the problems of integer programming. Let us consider two of them.

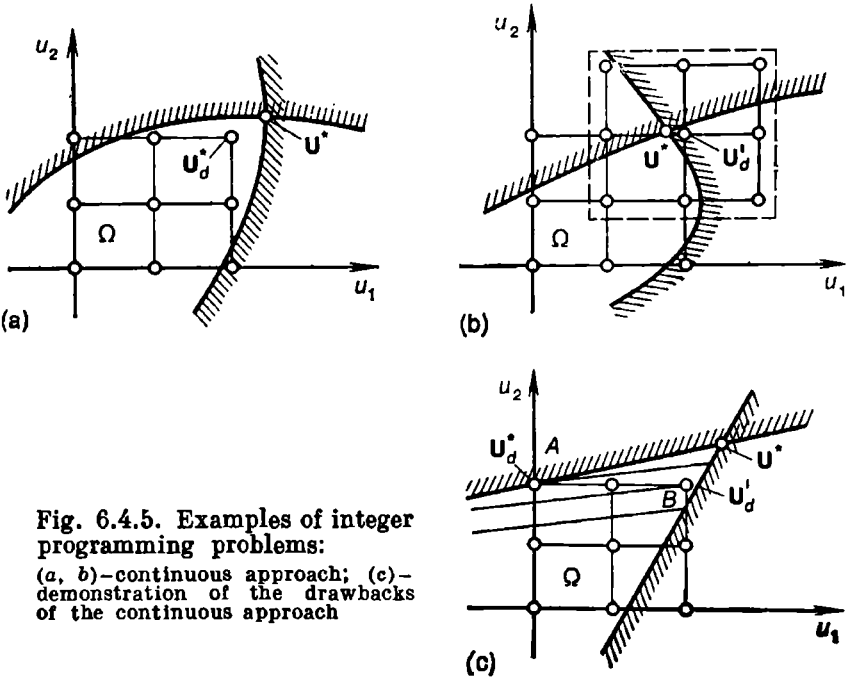

Fig. 6.4.5. Examples of integer programming problems:
(*a*, *b*)–continuous approach; (c)–demonstration of the drawbacks of the continuous approach

The *direct search* is based on the idea of a discrete random seeking in the process of which the travel is over the integer points of the admissible area Ω. Suppose $\mathbf{U}_N$ is an integer point within Ω. The process of the discrete random search implies random determination of a point in the integer vicinity of $\mathbf{U}_N$

$$\mathbf{U}_N^{\xi} = \mathbf{U}_N + \Xi,$$

where $\Xi = (\xi_1, \ldots, \xi_q)$ is a random integer vector with the components taking one out of the three values $\xi_i \in \{+1; 0; -1\}$ independently and with equal probability. If the value of the quality function Q in the point $\mathbf{U}_N^{\xi}$ proves to be less than in the basic point $\mathbf{U}_N$, then $\mathbf{U}_{N+1} = \mathbf{U}_N^{\xi}$. Otherwise, the step is unsuccessful and another random vector Ξ is to be found and the condition of decreasing the quality function verified again. Hence, the algorithm of the discrete random search may

be put down as

$$\mathbf{U}_{N+1} = \begin{cases} \mathbf{U}_N + \Xi, & \text{if } Q(\mathbf{U}_N + \Xi) < Q(\mathbf{U}_N), \\ \mathbf{U}_N, & \text{if } Q(\mathbf{U}_N + \Xi) \geqslant Q(\mathbf{U}_N). \end{cases} \tag{6.4.13}$$

Clearly, this implicates a random search in the integer vicinity of the basic point until another point is found with less value of the quality function. Let us note that since the set of the random vectors $\{\Xi\}$ is finite, it is advisable to exclude from further consideration at this stage the vectors Ξ already examined. Owing to such elimination of unnecessarily repeated operations, the efficiency of search is going to increase, and if a complete search does not produce a better point, that constitutes the simplest stop criterion.

For the problems of discrete convex programming, when $Q(\mathbf{U})$ and $\mathbf{H}(\mathbf{U})$ are convex functions, this method is convergent, i.e. inevitably leads to the solution $\mathbf{U}^*_d$ of the problem.

The *reverse search* is the travel from the point $\mathbf{U}'_d$, the nearest to $\mathbf{U}^*$ out of the admissible ones, i.e. belonging to the area Ω. Let us consider the discrete vicinity of this point consisting of $(3^q - 1)$ points. Finding all the admissible (i.e. belonging to Ω) points, we should select the one for which the value of the quality function Q is minimal. If that least value is more than for the basic point $\mathbf{U}'_d$, it is only natural then to assume $\mathbf{U}'_d$ to be the solution to the problem. But if it is less, we should examine the integer vicinity of that new point.

The reverse search helps solve the problems of discrete convex programming just as well as the direct one. But to carry it out it is necessary first to solve the continuous problem. The reverse search differs from the direct one in that it requires an exhaustive examination of the vicinity of every point supposed to be optimal. But there are essentially less such exhaustive searches in the reverse method than random searches in the direct one.

In a complicated case, it is recommended to use both methods and select the solution with the least value of the quality function.

6.5. Some Special Methods to Solve Variational Problems

It was shown in § 6.1 how the problem of synthesizing the dynamic object's control may be reduced to the problem of mathematical programming (6.1.8). It is sufficient for that to present the sought control $\mathbf{U}(t)$ as a set of discretes

$$\mathbf{U}_i = \mathbf{U}(t_i),\ i = 1, \ldots, k,$$

that may be determined by solving the problem of mathematical programming with regard to the vector parameters $\mathbf{U}_1, \ldots, \mathbf{U}_k$ (it is necessary to determine the total of kq parameters producing the optimal control).

However, this approach leads, as a rule, to an involved multiextremal problem of mathematical programming which is usually very difficult to solve. That is why, to solve the variational problems, it is often more convenient to apply some special methods, two of which are discussed below.

The maximum principle. This method is applied for dynamic objects whose model may be presented as a system of common differential equations

$$\dot{y}_i = f_i(y_1, \ldots, y_m),\ u(t)),\ i = 1, \ldots, m \tag{6.5.1}$$

with specified initial conditions at $t = 0$

$$y_i(0) = y_i^0 \quad (i = 1, \ldots, m)$$

and a scalar control. It is convenient to put down system (6.5.1) in the vector form

$$\dot{\mathbf{Y}} = \mathbf{F}(\mathbf{Y},\ u(t)),$$

where $\mathbf{Y}(t) = (y_1(t), \ldots, y_m(t))$ is the vector of the variables of the problem, and $\mathbf{F} = (f_1, \ldots, f_m)$ is the vector of the given functions defining the object's model.

It is necessary to solve the following problem: to take the object from the state $\mathbf{Y}_0 = (y_1(0), \ldots, y_m(0))$ into a specified state

$$\mathbf{Y}^* = (y^*_1, \ldots, y^*_m) \tag{6.5.2}$$

while the trajectory $\mathbf{Y}(t)$ and the control $u(t)$ should be such that to meet some specified constraints and the extremum objective, i.e. provide the minimum value of the specified functional

$$Q(u(t)) = \int_0^T f_0(\mathbf{Y}(t),\ u(t))\,dt, \tag{6.5.3}$$

where $f_0(\mathbf{Y}(t),\ u(t))$ is a given function. Various expenditures for the control process are commonly selected to form such a functional. At $f_0(\cdot) = 1$, we get $Q = T$, i.e. the time for taking the object from the state $\mathbf{Y}_0$ into $\mathbf{Y}^*$.

A typical example of a problem of this kind is the problem for the crane operator. It is required to move a heavy weight from one point to another in a minimal time. The operator handles triple control $u = \{-1;\ 0;\ 1\}$, i.e. the crane can travel in one ($u = -1$) or the other ($u = 1$) direction or stop ($u = 0$). An inexperienced operator usually sways the load and cannot 'calm' it for a long time. The optimal control $u^*(t)$ allows for stopping the load in a minimal time T precisely at the required point $\mathbf{Y}^*$.

Consequently, the given variational problem is put in the form

$$\int_0^T f_0(\mathbf{Y}(t),\ u(t))\,dt \to \min_{u(t)\in\Omega} \tag{6.5.4}$$

$$\Omega: \begin{cases} h_j(\mathbf{Y}(t),\ u(t)) \geqslant 0, & j = 1,\ \ldots,\ k_2, \\ \dot{y}_i = f_i(\mathbf{Y}(t),\ u(t)), & i = 1,\ \ldots,\ m, \\ y_i(0) = y^0{}_i, \end{cases}$$

where $h_j(\cdot)$ are some specified constraining functions of the inequality type. Let us note that there are no boundary conditions (6.5.2) in (6.5.4): they are going to appear later. So far we managed to formulate the problem of taking the object into a certain state $\mathbf{Y}(t)$ which is of little concern (so far!).

The maximum principle allows for reducing problem (6.5.4) to the problem of maximizing the so-called Hamil-

tonian function

$$\Gamma(u(t)) = \sum_{i=0}^{m} \psi_i(t) f_i(\mathbf{Y}(t), u(t)) \to \max_{u(t) \subset \Omega'} \tag{6.5.5}$$

with the constraints

$$\Omega' : h_j(Y(t), u(t)) \geqslant 0, \; i = 1, \ldots, k_2, \tag{6.5.6}$$

where the functions $\psi_i(t)$ are defined by solving the system of $(2m + 1)$ differential equations of the form

$$\dot{\psi}_0 = 0, \quad \dot{\psi}_i = -\sum_{j=0}^{m} \psi_j \frac{\partial f_j(\mathbf{Y}(t), u(t))}{\partial y_i} \tag{6.5.7}$$

$$\dot{y}_i = f_i(\mathbf{Y}(t), u(t)), \quad i = 1, \ldots, m$$

with the initial conditions

$$\begin{aligned} y_j(0) &= y_j^0, \qquad j = 1, \ldots, m, \\ \psi_i(0) &= c_i, \qquad i = 0, \ldots, m. \end{aligned} \tag{6.5.8}$$

The problem is interesting in that the maximization may be carried out in every time moment t since $\Gamma(u(t))$ is not a functional but a function and the maximization problem is unidimensional with regard to the parameter of control u. The minimized function at that certainly changes in time (see (6.5.7)) and in keeping with constraints (6.5.6), and the minimization should be carried out in every time moment t.

The solution of the problem at the given initial conditions (6.5.8) is not difficult. Indeed, it is reduced to the integration of the system of differential equations (6.5.7) and to solving unidimensional optimization problem (6.5.5) at every step of integration (for instance, by the golden section method described in the 'Supplement').

Now let us consider arriving at the specified point $\mathbf{Y}^*$. Obviously, the state $\mathbf{Y}(t)$ of the object at the moment t is determined by the initial conditions $\mathbf{C} = (c_0, \ldots, c_m)$ specified in (6.5.8), i.e. as a result of solving problem (6.5.5) we get $\mathbf{Y}(T)$ depending on the values of $\mathbf{C}$: $\mathbf{Y}(T) = f(\mathbf{C})$. Hence the problem to determine the initial con-

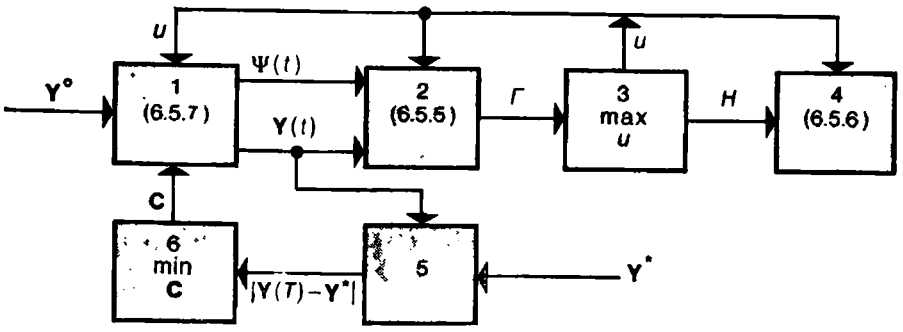

Fig. 6.5.1. Flow chart of computation with the maximum principle

ditions C is formulated clearly: they are to be selected so that $f(\mathbf{C}) = \mathbf{Y}^*$. But the dependence $f(\mathbf{C})$ is implicit, it is obtained algorithmically as a result of solving problem (6.5.5). Therefore it is only natural to reduce the procedure of finding C to solving the minimization problem

$$Q(\mathbf{C}) = |f(\mathbf{C}) - \mathbf{Y}^*| \to \min_{\mathbf{C}} \Rightarrow \mathbf{C}^*,$$

i.e. it is necessary to select such initial conditions C that the discrepancy function $Q(\mathbf{C})$ would be minimal (ideally—equal to zero). This problem should be solved by the seeking methods presented in the 'Supplement'. Finding the discrepancy Q for each new point C requires solving the problem (6.5.5). If we managed to find $\mathbf{C}^*$ such that $Q(\mathbf{C}^*) = 0$, we get the solution of problem (6.5.5) as the optimal control $u^*(t)$.

It is easy to notice that the procedure may be extended to the case of the multidimensional control $\mathbf{U}(t) = (u_1(t), \ldots, u_q(t))$ without difficulty. Minimization problem (6.5.5) becomes in this case multiparametrical, and it should be solved by one of the seeking methods (see the 'Supplement').

Let us consider the flow chart for the realization of calculations involving the maximum principle. It is shown in Fig. 6.5.1. Here *1* is the unit for solving the system of differential equations (6.5.7) whose input includes the control $u(t)$, the initial conditions C and $\mathbf{Y}^0 = (y_1^0, \ldots, y_m^0)$. The solution of this system of equations $\mathbf{Y}(t)$ and $\Psi(t) = (\psi_0(t), \ldots, \psi_m(t))$ is fed into unit *2* where

the Hamiltonian function is calculated according to (6.5.5) to be maximized by unit *3* with respect to u and observing the constraints (6.5.6) where $\mathbf{H} = (h_1, \ldots, h_{k_2})$ (this is the circuit of rapid optimization). There is the circuit of slow optimization solving the problem to minimize the discrepancy function $|\mathbf{Y}(t) - \mathbf{Y}^*|$ by means of changing the initial conditions **C**. For that the discrepancy between $\mathbf{Y}(t)$ and the objective state $\mathbf{Y}^*$ (unit *5*) is calculated at every step. Each step of minimization (unit *6*) is taken after each solution of system (6.5.7). It is not difficult in principle to realize such a system by applying modern electronic computers.

However, the quality function Q (**C**) of the problem of the external (slow) optimization is commonly of a multiextremal nature, which makes us employ the global seeking methods (presented in the 'Supplement') whose realization usually requires considerable consumption of the computer time. The consumption at that depends essentially on the dimension m of the object. This circumstance does not afford using the method to solve the problem of synthesizing the control of complex objects with a large dimension of m.

Dynamic programming. The method of dynamic programming is commonly applied for multiextremal variational problems whose solution may be reduced to a certain extent to organizing the exhaustive search. The method allows for putting in order and thus decreasing the number of exhaustive searches among all the admissible versions of control. It proves to be possible because the so-called 'optimum principle' is assumed to be real. Its essence is that the optimal behaviour of the object depends only on its current state and does not depend on the 'pre-history' of arriving at this current state. Thus, controlling an object, we should be concerned only with its state in the current moment because further control depends only on this state and not on its pre-history.

Naturally, the principle is applicable for the simple enough dynamic systems which are considered here. The 'simplicity' lies in that there is already a mathematical model of the controlled object allowing for good enough determination of the future behaviour of the object at this or that control.

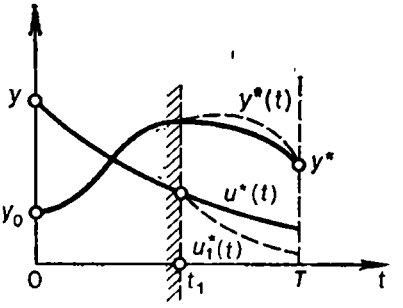

Fig. 6.5.2. To the optimum principle

Let us explain the gist of the method of dynamic programming with the example of the solution of the following variational problem which is a particular case of (6.5.4):

$$Q(u(t)) = \int_0^t f_0(y(t),\ u(t))\,dt \to \min_{u(t)\in\Omega} \tag{6.5.9}$$

where

$$\Omega: \begin{cases} \omega_1(t) \leqslant u(t) \leqslant \omega_2(t), \\ \dfrac{dy}{dt} = f_1(y,\ u(t)), \quad y(0) = y_0; \\ y(T) = y^*, \end{cases} \tag{6.5.10}$$

where the constraining functions $\omega_1(t)$ and $\omega_2(t)$ are somehow specified.

Problem (6.5.9) consists therefore in finding the optimal control $u^*(t)$ $(0 \leqslant t \leqslant T)$ for the first-order object (see (6.5.10)), whose behaviour $y(t)$ minimizes the given quality functional Q (6.5.9).

The solution to this problem, the optimal control $u^*(t)$ and optimal behaviour of the object $y^*(t)$ over the whole interval, is shown in Fig. 6.5.2. Now let us consider a moment t_1 and apply the optimum principle according to which the optimal control from the moment t_1 will depend only on the state at this moment, i.e. $y^*(t_1)$. This is obvious. It is natural to ask a question whether the optimal control proceeding from $t = 0$ and $t = t_1$ will be different within the interval $t_1 \leqslant t \leqslant T$, i.e. whether the optimal control and hence the object's behaviour will change as, for instance, is shown by the dashed line in Fig. 6.5.2. The answer to this not so trivial a

question follows from the optimum principle: no, it will not change!

Indeed, let us consider two realizations of the same methods to solve the initial problem, i.e. to find the optimal control within the interval $[0, T]$. Suppose in the first case we obtained an optimal solution $u^*(t)$ $(0 \leqslant t \leqslant T)$. Now let us divide the solution of the problem into two stages: $[0, t_1]$ and $[t_1, T]$ and solve the problem by the same method again. Suppose now the optimal solution is $u_1^*(t)$ $(t_1 \leqslant t \leqslant T)$ and it is different from $u_1^*(t)$ (see the dashed line in Fig. 6.5.2). But the result of solving the problem should not depend on the method of its solution (naturally, if both methods are correct). Therefore,

$$u_1^*(t) = u^*(t), \quad t_1 \leqslant t \leqslant T, \tag{6.5.11}$$

i.e. any part of $[t_1, T]$ of the optimal trajectory is *optimal*. This not quite evident consequence of the optimum principle underlies the computational procedure of dynamic programming. Let us note that it does not at all follow from (6.5.11) that any segment $[t_1, t_2]$ of the optimal trajectory is optimal from the viewpoint of the quality functional minimum over that time span. If $t_2 \neq T$, that is not so. Indeed, if the optimality over the segment $[t_1, t_2 < T]$ corresponded to the optimality over the whole interval $[0, T]$, that would mean that the control within $[t_1, t_2]$ would not influence the future behaviour of the object. That occurs in very uncommon cases that are, as a rule, of no interest since the problem degenerates at that.

Now let us discuss the method of dynamic programming. Let us first make problem (6.5.9) discrete (it will simplify the calculating algorithm), i.e. let us consider the control only in discrete time moments:

$$Q(\mathrm{U}) = \sum_{i=0}^{N-1} f_0(y_i, u_i)\,\Delta t_i \to \min_{u_i \in \Omega}, \tag{6.5.12}$$

$$\Omega: \begin{cases} \omega'_i \leqslant u_i \leqslant \omega''_i, \quad i = 0, \ldots, N, \\ y_{i+1} = y_i + f_1(y_i, u_i)\,\Delta t_i, \\ y_0 = y^0, \; y_N = y^*, \quad i = 0, \ldots, N-1, \end{cases}$$

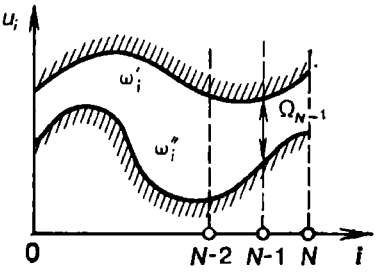

Fig. 6.5.3. Area of admissible controls

while $u_i = u(t_i)$; $y_i = y(t_i)$; $\Delta t_i = t_{i+1} - t_i$; $\omega'_i = \omega_1(t_i)$; $\omega''_i = \omega_2(t_i)$.

For simplicity let us consider the case of equidistant single-unit time moments, i.e. $\Delta t = 1$ $(i = 0, \ldots, N-1)$. We get

$$Q(\mathrm{U}) = \sum_{i=0}^{N-1} f_0(y_i, u_i) \to \min_{u_i \in \Omega}, \tag{6.5.13}$$

$$\Omega: \begin{cases} \omega'_i \leqslant u_i \leqslant \omega''_i, \\ y_{i+1} = y_i + f_1(y_i, u_i), \\ y_0 = y^0; \; y_N = y^*. \end{cases}$$

To solve this problem, let us make use of the optimum principle in form (6.5.11). We shall look for the solution 'from the rear' ($t = T$). The optimum principle guarantees that the optimal solution obtained over the segment $[t_1, T]$ will be a part of the solution of the prime problem for $[0, T]$.

Suppose the conditions Ω afford the control at the last $(N-1)$-th step within the limits Ω_{N-1}: $\omega'_{N-1} \leqslant u_{N-1} \leqslant \omega''_{N-1}$ (Fig. 6.5.3). The problem to be solved at that last step is to arrive at the point $y_N = y^*$, i.e. the problem to find the root of the equation

$$y^* = y_{N-1} + f_1(y_{N-1}, u_{N-1}) \Rightarrow u^*_{N-1}(y_{N-1}), \tag{6.5.14}$$

whose solution is the value of the control u^*_{N-1} dependent on y_{N-1} and independent from the value of the minimized functional Q at that. Indeed, optimization is no bother at the last step: it is necessary to meet the boundary condition $y_N = y^*$ at any price. The 'price'

depends on y_{N-1} and equals

$$\varphi_{N-1}(y_{N-1}) = f_0(y_{N-1}, u^*_{N-1}) \tag{6.5.15}$$

(recall that f_0 is the function of expenditures). Let us remember this.

Now moving from the end ($t = T$) to the beginning ($t = 0$), let us next consider the $(N-2)$-th step (the last but one) that should be taken from the hitherto unknown state y_{N-2}. The growth of the minimized criterion at that step equals $f_0(y_{N-2}, u_{N-2})$. In compliance with the optimum principle we should minimize not this value, but the total of the expenditures till the end of the process, i.e. the sum of the expenditures at this and the last step. But the expenditures for the last step are known (see (6.5.15)). In consequence, we should minimize the sum

$$f_0(y_{N-2}, u_{N-2}) + \varphi_{N-1}(y_{N-1}) \to \min_{u_{N-2} \in \Omega_{N-2}}, \tag{6.5.16}$$

where

$$y_{N-1} = y_{N-2} + f_1(y_{N-2}, u_{N-2}). \tag{6.5.17}$$

However, the initial conditions of this stage, i.e. the value y_{N-2}, are unknown. Therefore we have to 'store' the optimal control u^*_{N-2} for this step depending on the unknown value of y_{N-2}. For that we should solve problem (6.5.16) considering (6.5.17).

Let us denote the minimal value of the expenditures at the last two steps by the function

$$\varphi_{N-2}(y_{N-2}) = \min_{u_{N-2} \in \Omega_{N-2}} \{f_0(y_{N-2}, u_{N-2}) + \varphi_{N-1}(y_{N-2} + f_1(y_{N-2}, u_{N-2}))\}. \tag{6.5.18}$$

Proceeding similarly at the k-th from the end step, we get the expression of the total expenditures for the last k steps as

$$f_0(y_{N-k}, u_{N-k}) + \varphi_{N-(k-1)}(y_{N-(k-1)}).$$

Minimizing this expression with respect to u_{N-k} and regarding that

$$y_{N-(k-1)} = y_{N-k} + f_1(y_{N-k}, u_{N-k}),$$

we get $u^*_{N-k}(y_{N-k})$ and find as a result the function of the minimal expenditures

$$\varphi_{N-k}(y_{N-k}) = \min_{u_{N-k}\in\Omega_{N-k}} \{f_0(y_{N-k}, u_{N-k}) + \\ + \varphi_{N-(k-1)}(y_{N-k} + f_1(y_{N-k}, u_{N-k}))\}, \quad (6.5.19)$$

where Ω_{N-k} is the set of admissible controls at the k-th step from the end.

Consequently, moving 'from the rear', we obtain the following functions we shall need further on:

$$\varphi_1(y_1),\ u^*_1(y_1),\ \ldots,\ u^*_{N-1}(y_{N-1}) \quad (6.5.20)$$

that characterize the minimal expenditures (minimal value of the quality functional Q) while moving from the initial state y^0 into the objective one y^* and the dependence of the optimal control for each step on its own initial conditions.

Now the synthesis of the optimal control is very simple: it moves 'forward' from the beginning. The optimal control at the first step is found on the basis of the natural condition

$$f_0(y_0, u_0) + \varphi_1(y_0 + f_1(y_0, u_0)) \to \min_{u_0\in\Omega_0} \Rightarrow u^*_0,$$

the point

$$y^*_1 = y_0 + f_1(y_0, u^*_0)$$

being the point of the optimal trajectory at the first step. Then using the obtained functions (6.5.20) we shall have

$$u^*_1 = u^*_1(y^*_1),\ y^*_2 = y^*_1 + f_1(y^*_1, u^*_1) \text{ etc. to}$$

$$u^*_{N-1} = u^*_{N-1}(y^*_{N-1}).$$

Consequently, we have found the optimal control

$$\mathbf{U}^* = (u^*_0, u^*_1, \ldots, u^*_{N-1})$$

and the corresponding optimal trajectory

$$\mathbf{Y}^* = (y_0, y^*_1, \ldots, y^*_{N-1}, y^*).$$

Let us demonstrate the method with a simple example.

Example. It is necessary to find the optimal control $u(t)$ constrained as $|u(t)| \leqslant 1$ and taking the object $dy/dt = y + u$ from the state $y(0) = 0$ into the state $y(3) = 1$ ($T = 3$) with minimal expenditures calculated as the functional

$$Q(u(t)) = \int_0^3 (y^2 + u^2)\, dt,$$

i.e. it is necessary to solve the following variational problem:

$$Q = \int_0^3 (y^2 + u^2)\, dt \longrightarrow \min_{u(t) \in \Omega},$$

$$\Omega : \begin{cases} |u(t)| \leqslant 1, \\ \dfrac{dy}{dt} = y + u, \\ y(0) = 0;\ y(3) = 1. \end{cases}$$

Let us put down the problem in a discrete form assuming that $t_0 = 0$, $t_1 = 1$, $t_2 = 2$, $t_3 = 3$ and $y_i = y(t_i)$, $i = 0, 1, 2, 3$, i.e. dividing the integral $[0, T]$ into three parts:

$$Q = \sum_{i=0}^{2} (y^2_i + u^2_i) \longrightarrow \min_{U_i \in \Omega} \Rightarrow U^* = (u^*_0;\ u^*_1,\ u^*_2),$$

$$\Omega : \begin{cases} -1 \leqslant u_i \leqslant 1, \\ y_{i+1} = 2y_i + u_i \quad (i = 0, 1, 2),\ y_0 = 0,\ y_3 = 1. \end{cases}$$

It means that in this case $N = 3$, $\omega'_i = -1$, $\omega''_i = 1$,

$$f_0(y_i, u_i) = y^2_i + u^2_i, \quad f_1(y_i, u_i) = y_i + u_i,$$
$$y^0 = 0, \quad y^* = 1.$$

We shall start from the end ($t_2 = 2$). Let us find the control taking the object from the state y_2 into

$$y^* = 1 = 2y_2 + u^*_2.$$

At the last step (the second from the beginning) it equals

$$u^*_2(y_2) = 1 - 2y_2 \qquad (*)$$

while the minimal expenditures are

$$\varphi_2(y_2) = y^2_2 + [u^*_2(y_2)]^2 = 5y^2_2 - 4y_2 + 1.$$

Let us find the expenditures for the first step:

$$\varphi_1(y_1) = y^2_1 + u^2_1 + \varphi_2(y_2),$$

where $y_2 = 2y_1 + u_1$.

The total expenditures from $t = 1$ to $t = 3$ is

$$\varphi_1(y_1) = y^2_1 + u^2_1 + 5(2y_1 + u_1)^2 - 4(2y_1 + u_1) + 1.$$

Minimizing this function with respect to u_1, we find the dependence

$$u_1^* (y_1) = (1 - 5y_1)/3 \qquad (**)$$

and the minimal expenditures

$$\varphi_1 (y_1) = (13y_1^2 + 4y_1 + 1)/3. \qquad (***)$$

Now we can produce the optimal 'forward' control with the aid of the obtained functions:

Step 1. Substituting into (***) $y_1 = 2y_0 + u_0 = u_0$ (since $y_0 = 0$) and solving the problem

$$u^2{}_0 + \varphi_1 (u_0) \rightarrow \min_{|u_0| \leqslant 1},$$

we get $u_0^* = 1/8$.

Step 2. We find the first point of the optimal trajectory

$$y^*{}_1 = 2y_0 + u^*{}_0 = u^*{}_0 = 1/8.$$

Substituting this value into (**), we get the optimal control for the second step

$$u_1^* = (1 - 5/8)/3 = 1/8$$

and the point of the trajectory

$$y^*{}_2 = 2y^*{}_1 + u^*{}_1 = 3/8.$$

Step 3. And finally, we get from (*)

$$u^*{}_2 = 1 - 2y^*{}_2 = 1/4,$$

while

$$y^*{}_3 = 2y^*{}_2 + u^*{}_2 = 1 = y^*,$$

i.e. the object has been taken into the specified point.

Therefore, the optimal control is $U^* = (1/8, 1/8, 1/4)$ and the optimal trajectory $Y^* = (0, 1/8, 3/8, 1)$.

The expenditures in this case are

$$J^* = f_0 (y_0, u^*{}_0) + \varphi_1 (y^*{}_1) = 1/4.$$

Let us note that since the problem is simple, the minimization of the total expenditures at each step could be done analytically, and we could get the dependence of the optimal control on the state as a formula (see (**) in the example above). But this cannot be achieved in more involved cases, so it is necessary to present the dependence

$$u_i^* (y_i) \qquad (6.5.21)$$

in a tabular form, which naturally decreases the computational efficiency of the method.

Let us sum it up. As it was shown, the method of dynamic programming is a very efficient means to solve not

quite complicated variational problems. It allows for simple enough observing not only the constraints on the control $u_i \in \Omega_i$, but also the constraints on the optimal trajectory $y^*(t)$, i.e. it enables us to guide the solution through a specified 'corridor'. It is sufficient for that to solve the optimization problem at each step not for all the values of the coordinates, but only for those within the corridor at that step.

Consequently, the presented method of dynamic programming allows for solving variational problems of the form

$$Q(\mathbf{U}(t)) = \int_0^T f_0(\mathbf{Y}(t), \mathbf{U}(t))\, dt \to \min_{\mathbf{U}(t) \in \Omega}, \tag{6.5.22}$$

where

$$\Omega: \begin{cases} \mathbf{U}(t) \in \omega(t), \\ \mathbf{Y}(t) \in \xi(t), \\ \dfrac{d\mathbf{Y}}{dt} = \mathbf{F}(\mathbf{Y}, \mathbf{U}), \\ \mathbf{Y}(0) = \mathbf{Y}^0; \quad \mathbf{Y}(T) = \mathbf{Y}^*, \end{cases}$$

while $\mathbf{Y} = (y_1, \ldots, y_n)$; $\mathbf{U} = (u_1, \ldots, u_m)$ and specified are the areas $\omega(t)$ of the admissible control and $\xi(t)$ of the admissible trajectory—both of which may change with time—as well as the model $\mathbf{F}$ and the boundary conditions $\mathbf{Y}_0$ and $\mathbf{Y}^*$ of the problem.

One of the essential advantages of the method is the guarantee that the obtained solution is the global minimum of the functional Q. The point is that the functional is usually multiextremal in variation of the control $\mathbf{U}(t)$. Therefore the solution of the problem in case the optimum principle is invalid compels us to apply also the seeking methods for the global optimization (see the 'Supplement'), which always require appreciable consumption of the computer time. The presented approach eliminates the trouble of the multiextremal nature for this problem.

The disadvantages of the dynamic programming method should be noted as well. One of them is that it is practically unsuitable for solving the problems to control an object of the dimension more than $m = 2, 3$. The table

(6.5.21) in this case becomes multidimensional, which lowers appreciably the efficiency of the method since it brings about a sharp rise in the volume of the required memory (by about an order greater for each unit of dimension). This rules out applications of the method to control objects of a large dimension that commonly are complex objects of control.

To conclude this section, let us note that solving the problem to synthesize the optimal control of the dynamic object may be reduced to solving a variational problem whose complexity depends on the dimension and complexity of the object's model. The solution of the obtained problem to control the complex object is advisable to reduce to the solution of the problem of mathematical programming, and in some rare simple cases to the methods considered above and resting upon the idea of dynamic programming and the maximum principle described in this section. It should be borne in mind that there is an excellently elaborated theory for the discussed methods that (alas!) does not help solve the involved problems of large dimension existing in the practice of control of complex systems. Therefore, synthesizing the control of a complex dynamic object, one should apply the methods for solving multidimensional and multiextremal problems of mathematical programming, for instance, by the specially developed methods of random search (see the 'Supplement'). A good theory in solving this kind of problems is but a compass, and not an instrument or a tool. We have to solve the problems applying the apparatus of optimum seeking methods.

Comments on References to Chapter 6

The problem to synthesize control when the objective is specified and the object model is available is a problem of applied mathematics since it may be reduced to the solution of a problem of mathematical programming (that was shown in § 6.1) for which the required computational diagrams have been elaborated.

The classification of problems of mathematical programming proposed in § 6.2 is naturally incomplete, as is any classification. Its aim is to orient the reader with respect to the certain specificity of various types of problems of mathematical programming, the specificity being used in solving this kind of problems. More complete classifications of problems of mathematical programming can be found in the books [1-5].

There are many methods to solve the mathematical programming problems. Some of them were considered in § 6.4. All of them make use of various specific features of the functions Q, H, and G. The number of textbooks on mathematical programming is probably even too large. To make the first acquaintance with it, it is advisable for the reader to take [6-8].

Algorithms for solving the problems of linear programming are well described in the textbooks [1-6, 8-19, 22-24]. Problems of convex programming are analyzed and solved in the books [9, 15, 20, 25]. Methods to solve problems of integer programming are dealt with in more detail in the relevant chapters of the books [1-3, 13, 21, 24-28]. Besides, there are problems of geometrical programming, to make acquaintance with which the reader may take [29, 30]. Solution of conditional problems is described well in [33].

The existing methods to solve variational problems tackled with in § 6.5 are cumbersome in their algorithms and cover only a narrow class of the solved problems. Therefore the universal approach to their solution is still to reduce them to an appropriate problem of mathematical programming [31]. However, in some particular cases the methods presented in this book may produce appreciable effect. These methods are introduced in more detail in the textbooks [1-5, 7, 9, 12, 14, 32].

References

1. Kuzin, L.T. *Fundamentals of Cybernetics.* Energiya, Moscow, v. 1. *Mathematical Foundations of Cybernetics*, 1973 (in Russian).
2. Raskin, L.G. *Analysis of Complex Systems and Elements of the Theory of Control.* Sov. radio, Moscow, 1976 (in Russian).
3. Zaichenko, Yu.P. *Investigation of Operations.* Vishcha shkola, Kiev, 1975 (in Russian).
4. Pupkov, K.A. (ed.). *Foundations of Cybernetics*: *Mathematical Foundations of Cybernetics.* Vysshaya shkola, Moscow, 1974 (in Russian).
5. Intriligator, M. *Mathematical Optimization and Economic Theory.* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1971.
6. Gass, S.I. *An Illustrated Guide to Linear Programming.* McGraw-Hill Book Co., New York, 1970.
7. Lee, T.H., Adams, G.E., Gaines, W.M. *Computer Process Control*: *Modeling and Optimization.* Wiley, Inc., New York, 1968.
8. Kaufmann, A. *Conseiller scientifique à la compagnie des machines.* Bull. Dunod, Paris, 1962.
9. Gabasov, R., Kirillova, F.M. *Methods of Optimization.* Belorussian State Univ., Minsk, 1975 (in Russian).
10. Yudin, D.B., Golshtein, E.G. *Linear Programming*: *Theory, Methods and Applications.* Nauka, Moscow, 1969 (in Russian).
11. Gass, S.I. *Linear Programming. Methods and Applications.* McGraw-Hill Book Co., Inc., New York, 1958.

12. Korshunov, Yu.M. *Mathematical Foundations of Cybernetics.* Energiya, Moscow, 1972 (in Russian).
13. Lyashenko, I.N. (ed.). *Linear and Non-Linear Programming.* Vishcha shkola, Kiev, 1975 (in Russian).
14. Wagner, H. M. *Principles of Operations Research with Application to Managerial Decisions.* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1969, v. 1.
15. Zukhovitskii, S.I., Avdeeva, L.I. *Linear and Convex Programming.* Nauka, Moscow, 1967 (in Russian).
16. Romakin, M.I. *Elements of Linear Algebra and Linear Programming.* Vysshaya shkola, Moscow, 1963 (in Russian).
17. Karpelevich, F.I., Sadovskii, L.E. *Elements of Linear Algebra and Linear Programming.* Fizmatgiz, Moscow, 1963 (in Russian).
18. Karlin, S. *Mathematical Methods and Theory in Games, Programming, and Economics.* Pergamon Press, London, 1959.
19. Basov, A.S. *Linear Programming in Technical and Economical Problems.* Nauka, Moscow, 1964 (in Russian).
20. Fiacco, A.V., McCormic, G.P. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* Wiley, Inc., New York, 1968.
21. Korbut, A.A., Finkelshtein, Yu.Yu. *Discrete Programming.* Nauka, Moscow, 1969 (in Russian).
22. Yudin, D.B., Goldshtein, E.G. *Linear Programming.* Fizmatgiz, Moscow, 1963 (in Russian).
23. Gale, D. *The Theory of Linear Economic Models.* McGraw-Hill Book Co., Inc., New York, 1960.
24. Mukhacheva, E.A., Rubinshtein, G.Sh. *Mathematical Programming.* Nauka, Novosibirsk, 1977 (in Russian).
25. Saaty, T.L. *Optimization in Integers and Related Extremal Problems.* McGraw-Hill Book Co., New York, 1970.
26. Kovalev, M.M. *Discrete Optimization (Integer Programming).* Belorussian State Univ., Minsk, 1977 (in Russian).
27. Likhtenshtein, V.E. *Discreteness and Randomness in Economico-Mathematical Problems.* Nauka, Moscow, 1973 (in Russian).
28. Likhtenshtein, V.E. *Models of Discrete Programming.* Nauka, Moscow, 1971 (in Russian).
29. Zener, C. *Engineering Design by Geometrical Programming.* Wiley, Inc., New York, 1971.
30. Duffin, R.J., Peterson, E.J., Zener, C. *Geometric Programming—Theory and Application.* Wiley, Inc., New York, 1967.
31. Tabak, D., Kuo, B.C. *Optimal Control by Mathematical Programming.* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1971.
32. Moiseev, N.N. *Numerical Methods in the Theory of Optimal Systems.* Nauka, Moscow, 1971 (in Russian).
33. Gill, P.E., Murrey, W. (Eds.). *Numerical Methods for Constrained Optimization.* Academic Press, London, 1974.

Chapter 7 Realization of Control

Having solved the problem of synthesizing the optimal control $\mathbf{U}^*(t)$ considered in the preceding chapter, we can tackle the stage of control. It means that the controlled inputs $\mathbf{U}$ of the object should be taken into the state prescribed by the adopted decision $\mathbf{U}^*(t)$. The problem is solved very easily for simple objects of control in the automatic mode—by means of following the instruction (program) $\mathbf{U}^*(t)$ with the aid of appropriate servomechanisms automating the change in the object's controlled inputs $\mathbf{U}$. The control device in the 'manual' mode plays the part of an adviser indicating the positions the handles for the object's control should be taken into, while the realization of control itself is carried out by the personnel maintaining the object. There is no problem here—naturally, if there is no sign of trouble in the control device. The problems arise only in the control of complex objects. Let us consider them.

7.1. Formulation of the Problem

The problem to realize control appears when some important factors affecting the achievement of the objectives $\mathbf{Z}^*$ in the object were not taken into account in synthesizing the realized control $\mathbf{U}^*$. These factors, for instance, include the state $\mathbf{X}$ of the medium, the unobserved factors $\mathbf{E}$, and the deviation of the object $\mathbf{F}^o$ from its model $\mathbf{F}$. The reasons for the uncertainty are:

—the limited nature of any prognosis for the state of the medium. The point is that the control $\mathbf{U}^*$ was synthesized for a predicted state $\mathbf{X}$ which naturally and unavoidably differs from the real;

—the influence of unobserved factors $\mathbf{E}$ changing the situation shaping in the process of control both in the object and in the medium;

—the deviation of the object F^o from its model F. This deviation may be dual.

On the one hand, it may be caused by the drift of the object's parameters during the period from the moment of observing its state Y till the moment of realizing the control. On the other hand, it was intentionally set into the model F. For the sake of the model's simplicity, some information on the object is purposefully ignored, though it could be gained and used in the model's synthesis. The idea is clear: since in the process of realization of the control U^* there are always some unforeseen circumstances, the said deviation may be considered together with them so that the control might be corrected with due regard for all the concrete circumstances. That is why the realization of the control U^* is inevitably linked to its correction that takes into account all the conditions which have taken shape by the current moment. The new control has the form

$$(U^*)' = U^* + \delta U^*,$$

where δU^* is the correction of the control accounting for the above-mentioned unforeseen changes in the medium and the object and the methods to take them into account in the correction of the control.

7.2. Consideration of the Effects of the Medium

The static object. Suppose the optimal control U^* was synthesized for a certain predicted state of the medium $\hat{X}$. But in reality not the forecast $\hat{X}$, but $X \neq \hat{X}$ took place. The data on X allows for the correction of the control if it is known how the object should respond to the control U. Suppose the response is Y^*. Then the true response $Y = F(X, U^*)$ will differ from the required Y^* by the value of

$$\delta Y^* = F(X, U^*) - Y^*.$$

Here F is a function (since the model under consideration is that of the static object).

Obviously, the correction of control $\delta\mathbf{U}$ should minimize this value, i.e. for the synthesis of the correction $\delta\mathbf{U}$ it is necessary to solve the minimization problem

$$Q(\delta\mathbf{U}) = |\mathbf{F}(\mathbf{X},\ \mathbf{U}^* + \delta\mathbf{U}) - \mathbf{Y}^*| \to \min_{\delta\mathbf{U}\in R'}, \qquad (7.2.1)$$

where R' are the resources allocated for the correction.

Let us make use of the smallness of both $\delta\mathbf{U}$ and $\delta\mathbf{X}$, the deviation of the real state of the medium $\mathbf{X}$ from the predicted $\hat{\mathbf{X}}$: $\delta\mathbf{X} = \mathbf{X} - \hat{\mathbf{X}}$. Then we may put down the minimized function $Q(\cdot)$ as a linear decomposition

$$Q(\delta\mathbf{U}) = \left|\mathbf{F}(\hat{\mathbf{X}},\ \mathbf{U}^*) + \frac{\partial\mathbf{F}}{\partial\mathbf{X}}\delta\mathbf{X} + \frac{\partial\mathbf{F}}{\partial\mathbf{U}}\delta\mathbf{U} - \mathbf{Y}^*\right|$$

and supposing, for instance, that $n = m = q = 1$ and the resources R' are unlimited, we may get from the condition $Q(\delta\mathbf{U}) = 0$ the explicit form of the correction

$$\delta u = -\left(\frac{\partial f}{\partial u}\right)^{-1}\left[f(\hat{x},\ u^*) + \frac{\partial f}{\partial x}\delta x - y^*\right],$$

where the values of the derivatives of the function $f(\cdot,\ \cdot)$ are calculated for u^* and $\hat{x}$.

The dynamic object. Correction of the control in the process of its realization over the dynamic object requires to minimize the discrepancy between the desired change of the output of the object $\mathbf{Y}^*(t)$ in the process of control and what occurs in the medium $\mathbf{X}(t) \neq \hat{\mathbf{X}}(t)$, where $\hat{\mathbf{X}}(t)$ is the prediction of the medium behaviour. The problem has the form

$$\int_0^T [\mathbf{F}(\mathbf{X}(t),\ \mathbf{U}^*(t) + \delta\mathbf{U}(t)) - \mathbf{Y}^*(t)]^2\,dt \to \min_{\delta\mathbf{U}(t)\in R'},$$

where $\mathbf{F}$ is the operator of the dynamic model. However, it is necessary here to foresee the behaviour of the medium $\mathbf{X}(t)$, which may be achieved only by means of predicting further than $\hat{\mathbf{X}}(t)$, with whose aid the control $\mathbf{U}^*(t)$ was determined, and therefore more accurately.

7.3. Consideration of the Activity of the Object

In case the object proved to be active (for instance, an institution, workshop, factory, etc.) and this circumstance was not taken into account at the stage of the control synthesis, it cannot be avoided at the stage of the control realization. Let us show with a simple example the problems arising here.

The point is that the state $\mathbf{Y}$ of the active object depends not only on the medium $\mathbf{X}$ and control $\mathbf{U}$, but also on some of its internal parameters $\mathbf{W} = (w_1, \ldots, w_l)$, which the object may change at its own discretion in compliance with its own objectives $\mathbf{Z}^*_0$ (that is the essence of the object's activity). These objectives are seldom the same as the objectives $\mathbf{Z}^*$ of control, which brings about the troubles in the control over such an active object.

Consequently, the state of the active object may be presented as $\mathbf{Y} = \mathbf{F}_0(\mathbf{X}, \mathbf{U}, \mathbf{W})$, where $\mathbf{F}_0$ is the operator of the object.

Suppose the objectives $\mathbf{Z}^*_0$ and $\mathbf{Z}$ are extremal. Then the problem to be solved in synthesizing the control $\mathbf{U}$ has the form

$$\mathbf{Z}^*: \; Q_{\mathbf{u}}(\mathbf{U}, \mathbf{W}) \to \min_{\mathbf{U} \in R_{\mathbf{u}}} \Rightarrow \mathbf{U}^*_{\mathbf{w}}, \tag{7.3.1}$$

where $Q_{\mathbf{u}}(\cdot, \cdot)$ is the objective function of the control; $R_{\mathbf{u}}$ is the resource allocated for the control; $\mathbf{U}^*_{\mathbf{w}}$ is the taken decision that naturally depends on the values of the parameters $\mathbf{W}$ the object will choose. (The state of the medium is neglected for the notation simplicity's sake, i.e. we consider the medium to be stable.) But the object, being active and striving to reach its own objectives $\mathbf{Z}^*_0$, solves

$$\mathbf{Z}^*_0: \; Q_{\mathbf{w}}(\mathbf{U}, \mathbf{W}) \to \min_{\mathbf{W} \in R_{\mathbf{w}}} \Rightarrow \mathbf{W}^*_{\mathbf{u}}, \tag{7.3.2}$$

where $Q_{\mathbf{w}}$ is the objective function of the object; $R_{\mathbf{w}}$ is the resource at the object's disposal; $\mathbf{W}^*_{\mathbf{u}}$ is the decision made by the object and depending on $\mathbf{U}$. (Clearly, both decisions $\mathbf{U}^*_{\mathbf{w}}$ and $\mathbf{W}^*_{\mathbf{u}}$ depend on each other. This dependence may first of all lead to instability of the situation,

which is determined by the operator $\mathbf{F}_0$ of the object and the objective functions Q_{u*} and Q_{w*}, but that is not the main thing.)

Suppose the solutions $\mathbf{U}^*_w$ and $\mathbf{W}^*_u$ are stable. It is natural to ask a question: is it nonetheless possible to improve the current situation, i.e. to decrease the value of the objective function Q_u at the adopted decisions $\mathbf{U}^*_w$ and $\mathbf{W}^*_u$? It proves to be possible. It is necessary for that to consider the control as a process of interaction between two equal partners, i.e. take into account that the opposite party also tries to reach the extremal objectives. When both parties (the object and the control device) allow for this, it enables us to produce such a compromising control that should provide both parties with more favourable conditions than do $\mathbf{U}^*_w$ and $\mathbf{W}^*_u$ obtained in solving problems (7.3.1) and (7.3.2) with no allowance for the activity of the parties.

Let us demonstrate this with a simple example. Suppose

$$Q_w = u^2 + w^2, \quad Q_u = (u-1)^2 + (w-1)^2.$$

The 'relief' of both functions is shown in Fig. 7.3.1*a*. It consists of concentric circumferences with the centres (0, 0) and (1, 1). The solution of problems (7.3.1) and (7.3.2)

$$Q_u = (u-1)^2 + (w-1)^2 \longrightarrow \min_u \Rightarrow u^* = 1,$$

$$Q_w = u^2 + w^2 \longrightarrow \min_w \Rightarrow w^* = 0$$

leads to the stable point $\mathbf{B} = (1, 0)$ providing the partners with the single-unit values of their objective functions:

$$Q_u\,(1,\ 0) = Q_w\,(1,\ 0) = 1. \qquad (*)$$

That is all what one can reach with no consideration of a reasonable approach on the side of the partners. But as it can easily be seen in Fig. 7.3.1*a*, there is an area in the plane that can provide both parties with less values of their objective functions (it is cross-hatched in the figure). Since both objective functions are axially symmetrical, it is possible to consider their behaviour only along the straight line OA connecting the two objectives. This section is demonstrated in Fig. 7.3.1*b*. It can be well seen that the cross-hatched area is more attractive for both parties than the point **B**. Thus, if they agree to keep in the compromising point $C = (1/2,\ 1/2)$, then their quality indices are going to be $Q_u\,(1/2,\ 1/2) = Q_w\,(1/2,\ 1/2) = 1/2$, i.e. both parties have a two-fold gain from such a solution as compared to the point **B**. The

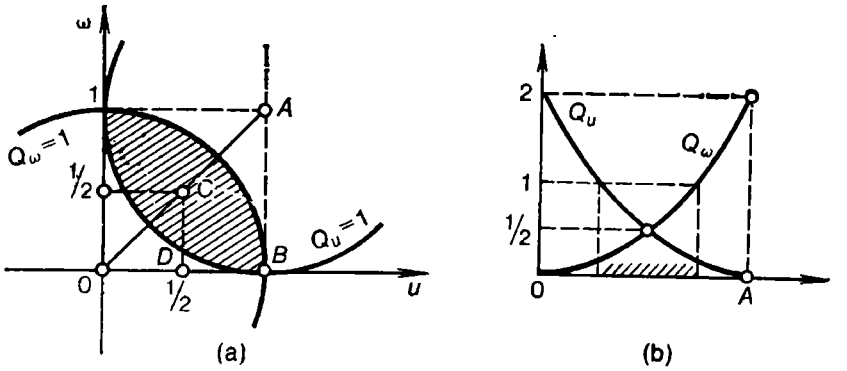

Fig. 7.3.1. Example of the 'relief' of the objective functions (*a*) and the quality indices along the AO axis (*b*)

obtained effect depends on the structure of the objective functions of both parties. What is then there to be done to realize this effect?

It is sufficient in this concrete case for each party to convey to the other party the value of its objective function and to reach an agreement to minimize, each by its own means, the sums

$$Q_w(u,\ w)+Q_u(u,\ w)\xrightarrow[u]{}\min \Rightarrow u^*=1/2,$$

$$Q_w(u,\ w)+Q_u(u,\ w)\xrightarrow[w]{}\min \Rightarrow w^*=1/2.$$

However, the obtained solution is unstable. Indeed, suffice it for any of the parties (for instance, the object) to violate the agreement and begin controlling again in compliance with its own objective function (7.3.2), the situation moves to the point $D = (1/2,\ 0)$, where the object has $Q_w\ (1/2,\ 0) = 1/4$, i.e. this situation for the object is by 1/4 better than the agreed. But the value of the objective function for the control device worsens by 3/4: $Q_u\ (1/2,\ 0) = 5/4$, and it has a right to come over to the control according to its own criterion Q_u, which pulls the system back into the stable point $\mathbf{B} = (1,\ 0)$. So the agreement is an unstable but useful state of the interaction between the two parties in the process of control.

Consequently, realizing the control of the active object, we should necessarily take into account its interests and stipulate the relations to meet not only the objectives of control, but the object's objectives as well which should provide for the efficient realization of the objectives $\mathbf{Z}^*$.

Comments on References to Chapter 7

The problem to realize control could be subdivided into two: prompt correction of control in connection with the emergence of a new situation in the medium and the object, and the control

realization proper. The second problem is related to the change in the position of the organs to control the object $u_1, \ldots, u_r$ by the actuating mechanisms in compliance with the optimal program $U^*(t)$ obtained at the stage of the control synthesis. The automatic following the program is carried out with the aid of the well developed methods of the theory of automatic regulation. There are many good textbooks in this fields; let us note some of them [1-10].

The object's activity considered in § 7.3, strictly speaking, should be regarded already at the stage of its model's synthesis. That means that it is necessary to know beforehand both the objectives of the active object and its possibilities to realize them. The formalization of these data results in a common model $F(X, U)$ whose synthesis was considered in Chapters 3-5. Some of the methods to formalize this kind of data are discussed in [11].

But if the subject has no such data at his disposal, then he discovers the result of the object's activity immediately at the stage of the control realization, since the object does not prove to be in the state the subject counted on, considering its model of the object. This is why the subject at the stage of the control realization has to 'come into contact' with the object, i.e. agree on the 'rules of the game' by whose virtue the object is allowed certain rights. The subject must also limit himself to certain agreed commitments with regard to the object. Now the account of these commitments forms a new objective within whose limits the optimal control should be produced. This intricate and conflicting situation is tackled with by the theory of games, in particular its new branch related to non-antagonistic games. One can introduce oneself into this new trend with the brochure [12] and the book [13].

An active object always includes a human being or a collective. Such an object is sometimes called ergatic. The specificity of the behaviour of ergatic systems is considered in the monograph [14]. The characteristics of the behaviour of the human being as the object of control are researched on in the publications [15, 16].

References

1. Netushil, A.V. (ed.). *The Theory of Automatic Regulation.* Vysshaya shkola, Moscow, 1976 (in Russian).
2. Netushil, A.V. (ed.). *The Theory of Automatic Regulation.* Vysshaya shkola, Moscow, 1972, Part II (in Russian).
3. Voronov, A.A. *Fundamentals of Automatic Regulation.* Energiya, Moscow, 1965. Part I. *Linear Systems of Regulation of One Value* (in Russian).
4. Voronov, A. A. *Fundamentals of Automatic Regulation.* Energiya, Moscow, 1966. Part II. *Special Linear and Non-Linear Systems of Automatic Regulation of One Value* (in Russian).
5. Voronov, A. A. *Fundamentals of Automatic Regulation.* Energiya, Leningrad, 1970. Part III. *Optimal Multiply-Connected and Adaptive Systems* (in Russian).

6. Pugachev, V.S. (ed.). *Fundamentals of Automatic Regulation.* Nauka, Moscow, 1968 (in Russian).
7. Saposhkov, R.A. *Fundamentals of Technical Cybernetics.* Vysshaya shkola, Moscow, 1970 (in Russian).
8. Yegorov, K.V. *Fundamentals of the Theory of Automatic Regulation.* Energiya, Moscow, 1967 (in Russian).
9. Yurevich, E.I. *The Theory of Automatic Regulation.* Energiya, Leningrad, 1969 (in Russian).
10. Besekerskii, V.A., Popov, E.P. *The Theory of the Systems of Automatic Regulation.* Nauka, Moscow, 1966 (in Russian).
11. Burkov, V.N. *Fundamentals of the Mathematical Theory of Active Systems.* Nauka, Moscow, 1977 (in Russian).
12. Vatel, I.A., Yereshko, F.I. *The Mathematics of Confict and Cooperation.* Znaniye, Moscow, 1973 (in Russian).
13. Germeyer, Yu.B. *Games with Non-Opposite Interests.* Nauka, Moscow, 1976 (in Russian).
14. Pavlov, V.V. *The Beginning of the Theory of Ergatic Systems.* Naukova dumka, Kiev, 1975 (in Russian).
15. Ackoff, R.L., Emery, F.E. *On Purposeful Systems.* Aldine & Atherton, Chicago, New York, 1972.
16. Lefevr, V.A. *Conflicting Structures.* Sov. radio, Moscow, 1973 (in Russian).

Chapter 8 Correction of the Control System

The necessity of correcting the control system arises in connection with the inconsistency of the object's response $\mathbf{Y}^o = \mathbf{F}^o(\mathbf{X}, \mathbf{U})$ and its model's response $\mathbf{Y} = \mathbf{F}(\mathbf{X}, \mathbf{U})$ as the result of the realization of the control $\mathbf{U}$ in the situation $\mathbf{X}$, i.e.

$$\mathbf{Y}^o \neq \mathbf{Y}. \qquad (8.0.1)$$

In case there are disturbances in the measurements of the object's states or the nature of its behaviour is stochastic, the mean value of $\mathbf{Y}^o$ is taken as true. When there is no statistical material to estimate the mean, the deviation of $\mathbf{Y}^o$ from $\mathbf{Y}$ should be reliable—for instance, beyond the 'three sigmas'.

In consequence, when inequality (8.0.1) is reliably true, a correction of the control system is needed. In compliance with Fig. I.4.1, this correction may concern the model, the object, and the objectives of control. In the first case, it may cover both the model's parameters and its structure. Evidently, the structure correction occurs less often than that of the parameters, which may be corrected at every step of the process of control.

If the model's correction does not succeed as expected, then the object's correction is needed, i.e. a change in the boundaries isolating it from the medium.

The process of realizing the objectives $\mathbf{Z}^*$ of the created control system is bound to bring about the correction of the initial set of objectives. The point is that the set of objectives reached by means of the realized control system does not, as a rule, coincide with the initial set of objectives. This is due to that the complex object evolves in a definite but unknown way, and it is only natural that not all the initial objectives can be realized in it. On the other hand, some new possibilities may be revealed to realize other objectives. All this causes the necessity of correcting the set of the objectives, i.e. of finding

out true possibilities of the given control system with regard to the given object.

Let us consider all the three cited kinds of the correction.

8.1. Correction of the Parameters of the Object's Model

Suppose the object's model has the form

$$\mathbf{Y} = \mathbf{F}(\mathbf{X}, \mathbf{U}, \mathbf{C}), \tag{8.1.1}$$

where $\mathbf{F}$ is, as usual, the model's structure, $\mathbf{C} = (c_1, \ldots, c_k)$ is its parameters. Both inputs $\mathbf{X}$ and $\mathbf{U}$ change in the process of control, which allows presenting the information on the operation of the object at the N-th step of control as the three measurements of the states

$$I_N = \langle \mathbf{X}_N, \mathbf{U}_N, \mathbf{Y}^o_N \rangle, \tag{8.1.2}$$

where $\mathbf{Y}^o_N = \mathbf{F}^o(\mathbf{X}_N, \mathbf{U}_N)$.

This three for the static object is a set of numbers, while for the dynamic one it is a set of functions

$$I_N(t) = \langle \mathbf{X}_N(t), \mathbf{U}_N(t), \mathbf{Y}^o_N(t) \rangle. \tag{8.1.3}$$

Let us consider the correction of the parameters of the models of static and dynamic objects in the case when the object's structure coincides with the model's structure, i.e.

$$\mathbf{F} = \mathbf{F}^o, \tag{8.1.4}$$

while the object's parameters change in a certain but unknown way, i.e.

$$\mathbf{C}^o = \mathbf{C}^o(t), \tag{8.1.5}$$

where $\mathbf{C}^o(t)$ is an unknown vector function of the drift of the object's parameters.

The problem to correct the parameters is to change the model's parameters $\mathbf{C}$ so that it would reflect the change in the object's parameters $\mathbf{C}^o(t)$, i.e.

$$\mathbf{C} \to \mathbf{C}^o(t). \tag{8.1.6}$$

Let us call the algorithm of solving this problem the algorithm of correction. It has to take into account the

information on the operation of the object and be of the recurrent nature, i.e.

$$\mathbf{C}_{N+1} = \mathbf{C}_N + \varphi(\mathbf{C}_N, I_N), \tag{8.1.7}$$

where φ is the algorithm of correction. Let us note that the recurrent nature of the correction algorithm is associated with the very set-up of the correction problem. This circumstance enables us to make use of the so-called method of self-adjusting model for the model's correction. A diagram to realize this method is shown in Fig. 8.1.1. The control system (it is shown by the dash line) realizes the objectives $\mathbf{Z}^*$ given to it from without. The information on the object's behaviour is compared with the response of the model $\mathbf{F}$ to the inputs $\mathbf{X}$ and $\mathbf{U}$, and the difference $\Delta\mathbf{Y}_N = \mathbf{Y}_N - \mathbf{Y}^o{}_N$ is fed into the correction unit that realizes the correction of the model's parameters:

$$\Delta\mathbf{C}_{N+1} = \varphi(\mathbf{C}_N, \mathbf{X}_N, \mathbf{U}_N, \Delta\mathbf{Y}_N). \tag{8.1.8}$$

Since the model $\mathbf{F}$ with the vector output $\mathbf{Y} = (y_1, \ldots, y_m)$ may be presented as m models $\mathbf{F} = (f_1, \ldots, f_m)$, the correction has to be carried out for each of the models

$$y_i = f_i(\mathbf{X}, \mathbf{U}, \mathbf{C}^i), \quad i = 1, \ldots, m,$$

where y_i is the output of the i-th model, f_i is its structure, and $\mathbf{C}^i$ is its parameters. As it was shown in § 3.3, we may consider the correction for only one model implying that the correction for the rest of them is going to occur similarly.

Therefore, further on we shall consider the correction of the model

$$y = f(\mathbf{X}, \mathbf{U}, \mathbf{C}) \tag{8.1.9}$$

with the scalar output y, where f is the known operator of the model.

It is easy to see that the process of correcting the model's parameters is very much like the process of the adaptive identification considered in Chapter 4. We shall use this similarity to synthesize the correction algorithms.

Let us consider the algorithm φ to correct the static and dynamic objects.

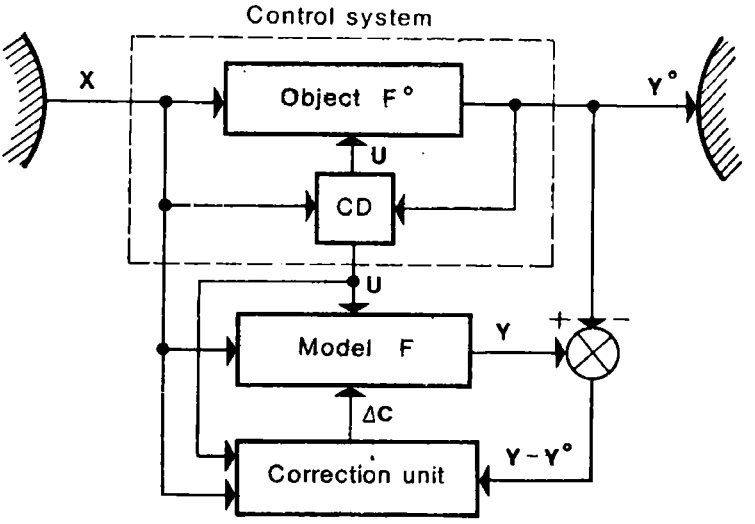

Fig. 8.1.1. Diagram of the model parameter correction by the method of self-adjusting model

The static object. The static object is characterized by the static structure described as the specified scalar function f (see (8.1.9)). Using the similarity to the adaptive identification, let us look for the correction algorithm among the algorithms of minimization of the local discrepancy between the outputs of the model and the object at the moment N:

$$q^2_N = (y_N - y^\circ_N)^2. \qquad (8.1.10)$$

Substituting the expression for the model's output (8.1.9), we get this discrepancy as a function of the model's parameters $\mathbf{C}_N$

$$q^2_N(\mathbf{C}_N) = (f(\mathbf{X}_N, \mathbf{U}_N, \mathbf{C}_N) - y^\circ_N)^2. \qquad (8.1.11)$$

Let us minimize this discrepancy by changing (correcting) the model's parameters, i.e. by solving the problem

$$q^2_N(\mathbf{C}_N + \Delta\mathbf{C}_{N+1}) \to \min_{\Delta\mathbf{C}_{N+1}}.$$

It is natural here to use the correction which would provide at the minimal change of $\mathbf{C}_N$ the greatest decrease in the discrepancy q^2. Such a correction is the gradient cor-

rection

$$\Delta \mathbf{C}_{N+1} = -\alpha \nabla_{\mathbf{C}} q^2{}_N (\mathbf{C}_N), \qquad (8.1.12)$$

where α is a positive parameter (it will be discussed below); $\nabla_{\mathbf{C}} q^2{}_N$ is the gradient with regard to $\mathbf{C}$ of discrepancy function (8.1.11) at $\mathbf{C} = \mathbf{C}_N$, the gradient being equal by definition to

$$\nabla_{\mathbf{C}} q^2{}_N = 2 q_N \nabla_{\mathbf{C}} f(\mathbf{X}_N, \mathbf{U}_N, \mathbf{C})\,|_{\mathbf{C}=\mathbf{C}_N}, \qquad (8.1.13)$$

where

$$\nabla_{\mathbf{C}} f(\mathbf{X}_N, \mathbf{U}_N, \mathbf{C}) = \left(\frac{\partial f(\mathbf{X}_N, \mathbf{U}_N, \mathbf{C})}{\partial c_1}, \ldots, \frac{\partial f(\mathbf{X}_N, \mathbf{U}_N, \mathbf{C})}{\partial c_k} \right).$$

These expressions are easily calculated if the structure of the model's function f is analytical. But if the function is very complicated or specified algorithmically as, for instance, in the case of the simulation model, then its derivatives may be estimated as follows:

$$\widehat{\frac{\partial f}{\partial c_i}} = \frac{1}{2g_i} [f(\mathbf{X}_N, \mathbf{U}_N, \mathbf{C}_N + \mathbf{e}_i g_i) - f(\mathbf{X}_N, \mathbf{U}_N, \mathbf{C}_N - \mathbf{e}_i g_i)], \qquad (8.1.14)$$

where g_i is the estimate basis; $\mathbf{e}_i$ is the i-th unit vector.

Now the expression for the correction may be put down in the form

$$\Delta \mathbf{C}_{N+1} = -2\alpha [f(\mathbf{X}_N, \mathbf{U}_N, \mathbf{C}_N) - y^o{}_N] \nabla_{\mathbf{C}} f(\mathbf{X}_N, \mathbf{U}_N, \mathbf{C})\,|_{\mathbf{C}=\mathbf{C}_N} \qquad (8.1.15)$$

Let us consider the problem to select the parameter α. If there are no disturbances, it should be set for the reasons of the best possible minimization of the discrepancy. And since its minimal value cannot be less than zero, the selection of α is related to the solution of the problem considered in Chapter 4:

$$q^2{}_N (\mathbf{C}_N + \Delta \mathbf{C}_{N+1}) \to \min_{\alpha} = 0 \Rightarrow \alpha^* \qquad (8.1.16)$$

(see Fig. 4.2.1). This is a problem of monoparametrical minimization that may be solved by any of the known search methods (e.g. by the golden section method described in the 'Supplement', § S.2).

But if the initial information on the behaviour of the object contains much 'noise', then the parameter α should

be less than α^*. Indeed, the value of the correction ΔC_{N+1} is proportional to the discrepancy $(y_N - y^o{}_N)$; if there are disturbances, the discrepancy is appreciable not because of the model's inaccuracy, but because of the disturbances superimposed in the process of measuring the state of the object. Therefore the parameter α should decrease the correction, and the greater the level of the disturbances, the more the decrease.

Consequently, the parameters of the static model are to be corrected according to formula (8.1.15), and the correction is linked to either calculation or estimation of the gradient of the model's output function over the corrected parameters (8.1.14).

The dynamic object. This is the case when f in (8.1.9) is an operator of a certain form while C is the vector of its parameters. Let us consider the discrete dynamic object whose operator has a finite difference form of the (4.3.4) kind. Its model, when the control is present, may be put down as

$$y_{N+1} = f(\mathbf{X}'_{N+1}, \mathbf{U}_{N+1}, \mathbf{Y}'_N, \mathbf{C}),$$

where $\mathbf{Y}'_N = (y_N, y_{N-1}, \ldots, y_{N-p})$; f is a given function of the four vector arguments $\mathbf{X}'$, $\mathbf{U}$, $\mathbf{Y}'$ and $\mathbf{C}$ specified at the stage of the structural synthesis.

Suppose that the response of the model y_N and the object $y^o{}_N$ do not coincide, i.e. $y_N \neq y^o{}_N$. This makes us start correcting the model's parameters $\mathbf{C}$ by means of minimizing the discrepancy function, i.e. solving the problem

$$q^2{}_N(\mathbf{C}) = [f(\mathbf{X}'_N, \mathbf{U}_N, \mathbf{Y}'_{N-1}, \mathbf{C}) - y^o{}_N]^2 \to \min_{\mathbf{C}}.$$

It is natural to use the gradient method here:

$$\mathbf{C}_{N+1} = \mathbf{C}_N - \alpha \nabla_{\mathbf{C}} q^2{}_N(\mathbf{C})\,|_{\mathbf{C}=\mathbf{C}_N},$$

where in order to estimate the gradient of the (8.1.13) kind, it is necessary to have the estimates of the derivatives similar to (8.1.14):

$$\frac{\partial f}{\partial c_i} = \frac{1}{2g_i}[f(\mathbf{X}'_N, \mathbf{U}_N, \mathbf{Y}'_{N-1}, \mathbf{C}_N + \mathbf{e}_i g_i) - f(\mathbf{X}'_N, \mathbf{U}_N, \mathbf{Y}'_{N-1}, \mathbf{C}_N - \mathbf{e}_i g_i)], \quad i = 1, \ldots, k.$$

The parameter α here is selected in the same way as for the static object.

In consequence, the main obstacle in correcting the object's parameters is the estimation of the gradient of the function of local discrepancy between the outputs of the object and the model with respect to the corrected parameters. However, the cited procedure is carried out only on the basis of the information acquired as a result of the control realized in the object, and it corrects the model just for this control. If the next stage requires other control taking into account some hitherto unknown properties of the object (as it happens in most cases), then the model will not be 'ready' for that. This compels to add to the optimal control $\mathbf{U}^*(t)$ the so-called probing addition $\delta\mathbf{U}(t)$ allowing for better correction of the model without distorting the optimal control very much. As a matter of fact, this is the experiment design in the process of control. The addition $\delta\mathbf{U}(t)$ should be determined so that the correction $\Delta\mathbf{C}$ of the model's parameters on its basis would provide a possibility to produce a more accurate model.

Clearly, the summarized control $\mathbf{U}^*(t) + \delta\mathbf{U}(t)$ has two aims: to reach the given objectives $\mathbf{Z}^*$ of the control and simultaneously to produce the object's model $\mathbf{F}$. Obviously, the control in that case should be optimal. Such a control with a double aim acquired the name of dual control. Its methods are related to the solution of the described problem, but they are very complicated and therefore will not be presented here. Their essence is the minimization of a certain function estimating the overall damage caused by not reaching the objective of control (because of the addition $\delta\mathbf{U}(t)$) and by the model's inaccuracy. The minimization of these losses makes it possible to determine the optimal dual control.

8.2. Correction of the Model's Structure

We shall consider the correction of the model's structure as a passage from one structure to another in the process of control.

Suppose it is known that any change in the object's structure is within a finite number of some given structures (functions or operators) $f^1, \ldots, f^k$ of the model. This means that at each time moment one of these models

$$y^i = f^i(\mathbf{X}, \mathbf{U}, \mathbf{C}^i), \quad i = 1, \ldots, k \qquad (8.2.1)$$

describes best the changing object's structure

$$y_t^o = f_t^o(\mathbf{X}, \mathbf{U}),$$

where $\mathbf{C}^i$ is the vector of the i-th model's parameters; f_t^o is the object's structure somehow changing with time t.

It is natural to select as the criterion of the structure's efficiency an averaged discrepancy between each model and the object estimated before correcting the parameters (otherwise the discrepancies are going to be zero or close to it). Suppose for the static object

$$Q_j^i = [f^i(\mathbf{X}, \mathbf{U}, \mathbf{C}_j^i) - y_j^o]^2$$

is the discrepancy of the i-th model at the j-th step of control before correcting its parameters $\mathbf{C}^i$. However, the obtained values of the discrepancies at the j-th step of control $Q_j^i, \ldots, Q_j^k$ speak very little of the efficiency of each of the k models since random factors of any kind influence these values. Therefore it is natural to introduce the efficiency indices for the structures in the form of weighted sums of these discrepancies

$$\bar{Q}^i{}_N = \sum_{j=1}^{N} g^{N-j} Q^i{}_j, \quad i = 1, \ldots, k, \qquad (8.2.2)$$

where $0 \leqslant g < 1$ is the index of 'forgetting' that 'depreciates', as it were, the data obtained at the previous step (the deeper 'into the past', the less important the measurements). At $g = 0$ we get $\bar{Q}_N^i = Q_N^i$, i.e. the weighted discrepancy is equal to the discrepancy at the last step since in this case the entire pre-history is 'forgotten' (recall that $0^0 = 1$).

It is possible to generalize the expression for the efficiency indices (8.2.2) replacing the weight coefficient g^{N-j} by the function $\varphi(j)$:

$$\bar{Q}^i{}_N = \sum_{j=1}^{N} \varphi(N-j) Q^i{}_j, \quad i = 1, \ldots, k.$$

Here $\varphi(\cdot)$ is a monotonically decreasing 'forgetting' function with the following properties:

$$\Xi: \varphi(i) > 0, \quad \varphi(0) = 1, \quad \lim_{i \to \infty} \varphi(i) = 0,$$

$$\varphi(i) \geqslant \varphi(i+1) \quad (i = 0, 1, \ldots). \tag{8.2.3}$$

It is rather easy now to work out the solving rule. Obviously, the most fitting is the model whose efficiency index is minimal. This means that at the N-th step the decision on the selection of the structure is made as follows:

$$f_N = f^l \text{ if } \bar{Q}^l{}_N = \min_{i=1, \ldots, k} \bar{Q}^i{}_N. \tag{8.2.4}$$

In consequence, the procedure of adapting the structure is reduced to estimation of the efficiency index at each step for all the competing structures and the selection of the structure whose efficiency index is the least. Clearly, to realize the procedure, it is necessary to store and constantly correct the parameters of the k models out of which only the best one is used to synthesize the control.

When there are disturbances, the selection of the 'forgetting' function $\varphi(\cdot)$ is especially important. The higher the level of disturbances in the object, the slower the decrease of the function should be (thus carrying out the filtration). However, on the other hand, when the object's structure changes intensely, the function should decrease rapidly enough to 'forget' the results of the preceding experiments sooner since they play but an insignificant role because of the fast change in the object's structure. Therefore, to synthesize the optimal solving rule (8.2.4), it is important to select the function $\varphi(\cdot)$ properly. How shall we do that?

It is necessary for that to estimate the probability P of making a wrong decision, i.e. of deciding to select an l-th structure while in fact the best is another structure. This probability depends on the function $\varphi(\cdot)$. Now the problem to optimize the solving rule (8.2.4) should be presented as the problem to minimize the functional of probability of the error:

$$P(\varphi(i)) \to \min_{\varphi(i) \in \Xi} \Rightarrow \varphi^*(i), \tag{8.2.5}$$

where Ξ is the set of admissible functions with properties (8.2.3); $\varphi^*(i)$ is the optimal 'forgetting' function.

To estimate the error probability P, it is necessary to know the intensity and nature of the object's structure drift, i.e. its stochastic characteristics.

8.3. Correction of the Controlled Object

The evolution of the complex object $\mathbf{F}^o$ in the process of control, the drift of its parameters and characteristics, the change in the medium, i.e. the set of states $\{\mathbf{X}\}$, as well as the change in the set of objectives $\{\mathbf{Z}^*\}$ may result in that despite the adaptation of the parameters and the object's model structure the control system would not be able to provide the efficient realization of all the given objectives to be reached in the object. This compels us to reconsider the interface between the object and the medium whose initial version was determined at the stage of isolating the object from the medium (see Chapter 2).

To correct the object in the process of control, it is necessary to have alternatives, i.e. versions of the object from which it would be possible to select in correction. Suppose the number of such alternatives is finite and they are

$$\mathbf{F}^{o1}, \mathbf{F}^{o2}, \ldots, \mathbf{F}^{og}. \qquad (8.3.1)$$

Each version of the object out of (8.3.1) should have its model which is synthesized by the methods considered above (see Chapters 3-5):

$$\mathbf{Y}^i = \mathbf{F}^i(\mathbf{X}^i, \mathbf{U}^i), \; i = 1, \ldots, g, \qquad (8.3.2)$$

where $\mathbf{F}^i$ is the operator of the i-th model of the object $\mathbf{F}^{oi}$; $\mathbf{X}^i$ are the observed, and $\mathbf{U}^i$ the controlled inputs of the object. The inputs differ both quantitatively and qualitatively. The outputs $\mathbf{Y}^i$ are qualitatively the same for all the objects but different in their values. Therefore it is only natural to estimate the efficiency of the object by the degree of its achieving the set objective $\mathbf{Z}^*$ of control. It is necessary to introduce for that a measure of non-achievement of the objective as a scalar function of

two variables

$$\nu = \nu (Z^*, \Psi (Y)), \tag{8.3.3}$$

where $\Psi (\cdot)$ is the transformation of the space of states $\{Y\}$ of the object into the objective space $\{Z\}$ (see (1.3.6)).

For the objectives of the equality type this function could be defined, for instance, as $\nu = | Z^* - \Psi (Y) |^2$. It is somewhat more involved to do that for the objectives (constraints) of the inequality type, and difficult though possible for the extremal objectives. In either event let us assume that function (8.3.3) is determined and there is a possibility for a quantitative estimation of the degree of realization of the given objective in the object.

Now suppose for the sake of simplicity that the objective Z^* and the state X of the medium do not change with time (this is the problem to stabilize the object in a specified state $Y^* = \Psi^{-1} (Z^*)$, where Ψ^{-1} is the inverse function of Ψ), and there are no disturbances. Then the solution of the problem to select the best object out of (8.3.1) is a consequtive exhaustive search and controlling each of the g competing objects, while storing the values

$$\nu^i = \nu (Z^*, \Psi (Y^i)), \quad i = 1, \ldots, g.$$

The procedure ends in the selection of such an object that would provide the least value of ν. This rule may be put down in a simple form

$$F^0 = F^{0l} \text{ if } \nu^l = \min_{i=1, \ldots, g} \nu^i. \tag{8.3.4}$$

But such an exhaustive search is unsuitable when the object's properties change rapidly since both 'bad' and 'good' objects are going to be taken equally (to be more accurate, almost equally) often. This is a case when it is advisable to resort to a randomized rule. It is necessary for that to estimate the probabilities of selecting the object

$$p^1, \ldots, p^g \left(\sum_{i=1}^{g} p^i = 1 \right), \tag{8.3.5}$$

where p^i is the probability of the i-th object's selection. These probabilities can be connected with the values of

the efficiency indices

$$p^i = f(\nu^i),\ i = 1, \ldots, g,$$

where f is the inversely proportional function of the form $f(\nu) = a/(\nu + b)$, the parameter a being a normalizing multiplier to meet condition (8.3.5) at each step while the value of the parameter b should be equal to the estimate of the mean efficiency for all objects

$$b = \frac{1}{g} \sum_{i=1}^{g} \nu^i$$

and therefore it should be adjusted in the process of control.

At the initial moment, $\nu^i = 0$ $(i = 1, \ldots, g)$. Further, in the process of running this or that version of the object, probabilities (8.3.5) will take the values

$$p^i = a/(\nu^i + b), \tag{8.3.6}$$

where ν^i is the efficiency of the last run of the i-th version of the object. Clearly, this provides for taking the best object more often, thus ensuring the object's correction, i.e. the adjustment of its boundaries to its changing properties.

Now let us generalize it and remove the constraints on the state **X** of the medium, the objective **Z***, and the presence of disturbances. These unpredictably changing factors make the result of estimating ν stochastic, compelling to resort to filtering procedures of the type

$$\bar{\nu}^i{}_{N+1} = (1-\alpha)\,\bar{\nu}^i{}_N + \alpha \nu^i{}_{N+1}, \tag{8.3.7}$$

where α is the forgetting index $(0 < \alpha \leqslant 1)$; $\bar{\nu}^i{}_{N+1}$ is the mean value of the efficiency of the i-th object's version at the $(N+1)$-th run; ν^i_{N+1} is the estimate of the efficiency of that object according to (8.3.3) at the $(N+1)$-th run.

The procedure of making decisions on selecting an object does not change in this case. It is randomized, and the probabilities (8.3.5) of the alternatives are equal to (8.3.6) where instead of ν^i, it is necessary to substitute (8.3.7). Obviously, this procedure with a larger 'forget-

ting' ($\alpha = 1$) degenerates into a procedure considered above, where the probabilities p^i are determined only by the last estimate of the efficiency v^i.

8.4. Correction of the Objectives of Control

Suppose Λ is the set of the objectives $\{\mathbf{Z}^*\}$ to be realized by the produced control system. This implies that it was produced so that for any objective $\mathbf{Z}^* \in \Lambda$ and the state of the medium $\mathbf{X} \in \mathbf{S}$ a control would be found to take the object into the state Y corresponding to the realization of the objective $\mathbf{Z}^*$. However, the approximate nature of the model, the drift of the object's characteristics, and the limited possibilities of the system to correct the control in the period of its synthesis and realization often result in that the set of realized objectives differs essentially from Λ. Besides, the evolution of the object and the change in the properties of the medium bring about the drift of the set of realized objectives. Obviously, the boundaries of that set should be of interest for the consumer of the control system since they determine its ultimate possibilities.

Let us consider the procedure to correct the boundaries of the set Λ in the maintenance process of the control system. First of all, let us find out what affects the realizability of the objective $\mathbf{Z}^*$. Its realizability depends on the nature of the objective itself and on the state X of the medium where the object of control is. Therefore, the pair

$$\mathbf{L} = \langle \mathbf{X}, \mathbf{Z}^* \rangle \qquad (8.4.1)$$

determines the success or failure of the control. Hence, the boundary of the realizability may be outlined with the aid of the solving function

$$R = R(\mathbf{L}), \qquad (8.4.2)$$

assuming two values (0 or 1). In case $R(\mathbf{L}) = 0$ this means the objective $\mathbf{Z}^*$ at the state X of the medium cannot be realized by the given control system. But if $R(\mathbf{L}) = 1$, then the objective $\mathbf{Z}^*$ can be realized by the given control system at the state X of the medium.

In consequence, the possibilities of the control system may be characterized by function (8.4.2). The evolution of the object and the medium affecting the realizability of the objectives should also change this solving function.

Two problems arise in this connection:

—how to produce such a solving function for the given control system;

—how to correct this function according to the change (evolution) of the object and the medium (and possibly, of the control system itself).

Let us begin with solving the first problem. It is necessary for that to have the initial information on the experience of the realization of certain control objectives at certain states of the medium, i.e. the so-called training series formed by the pairs of correspondence

$$\mathbf{L}_i \to R_i, \; i = 1, \ldots, N, \qquad (8.4.3)$$

where $\mathbf{L}_i$ is the i-th pair (8.4.1), and R_i is the result of the control system's operation with this pair (the objective $\mathbf{Z}_i^*$ with the state $\mathbf{X}_i$ of the medium is not reached at $R_i = 0$, and it is reached at $R_i = 1$). Using these data, it is necessary to obtain a related solving function R (**L**) assuming two values (0 or 1). This problem is called the pattern recognition problem, and its essence is producing such a solving rule (recognizing function) according to known training series (8.4.3).

There are many methods to produce solving rules for pattern recognition problems (see the comments on § 8.4). The so-called method of potential functions fits best for the solution of the given problem. Let us consider it.

First, let us generate the space {**L**} where each pair (8.3.1) is presented as a point. The coordinates of this space are the observed parameters of the medium **X** and the objective **Z***. Therefore, series (8.4.3) is presented in the space {**L**} as points of two kinds. Some of them ($R = 0$) reflect the negative experience of using the given control system, others ($R = 1$)—the positive experience.

The problem to synthesize the solving rule R (**L**) consists in outlining in the space {**L**} a surface dividing the

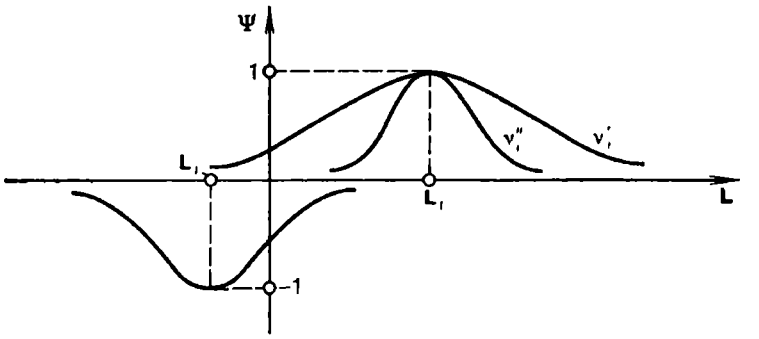

Fig. 8.4.1. Example of the potentials caused by positive (L_i) and negative (L_j) 'charges' ($\nu''_i > \nu'_i$)

points with the indices $R = 0$ and $R = 1$. Suppose each point $\mathbf{L}_i$ has a 'potential' in the space $\{\mathbf{L}\}$

$$\psi_i(\mathbf{L}) = \frac{q_i}{1 + \nu_i \rho(\mathbf{L}, \mathbf{L}_i)}, \tag{8.4.4}$$

where q_i is the 'charge' at the point $\mathbf{L}_i$ generating this potential:

$$q_i = \begin{cases} +1 \text{ if } R_i = 1, \\ -1 \text{ if } R_i = 0; \end{cases} \tag{8.4.5}$$

$\nu_i > 0$ is a certain constant of the potential attenuation; $\rho(\mathbf{L}, \mathbf{L}_i)$ is a function of the distance between the points $\mathbf{L}$ and $\mathbf{L}_i$, for instance,

$$\rho(\mathbf{L}, \mathbf{L}_i) = |\mathbf{L} - \mathbf{L}_i|. \tag{8.4.6}$$

Potentials (8.3.4) for different signs of the charge are shown in Fig. 8.4.1. The total potential

$$\Psi(\mathbf{L}) = \sum_{i=1}^{N} \psi_i(\mathbf{L}) \tag{8.4.7}$$

determines the sign of the charge in any other point of the space and hence the realizability of the control in that point

$$R(\mathbf{L}) = \begin{cases} 1 \text{ for } \Psi(\mathbf{L}) \geqslant 0 \\ 0 \text{ for } \Psi(\mathbf{L}) < 0 \end{cases} \tag{8.4.8}$$

or R (L) $= 1/2$ [1 + sign Ψ (L)], where 'sign' is the sign function

$$\operatorname{sign} z = \begin{cases} +1 \text{ for } z \geqslant 0, \\ -1 \text{ for } z < 0. \end{cases}$$

This is the sought solving function unambiguously indicating the realizability of the objectives of control according to observations of the realization of a certain number (N) of the objectives $\mathbf{Z}_i^*$ in concrete situations $\mathbf{X}_i$ by the given control system.

Now let us consider the correction of the solving function (8.4.8). The need in such a correction is linked to the 'ageing' of data (8.4.3), i.e. its inevitable depreciation. It is very easy to take that ageing into account by using the 'forgetting' function $\varphi(\cdot)$ introduced above (see (8.2.3)). It is sufficient for that to decrease the potential q_i as compared with (8.4.5) in the following way:

$$q_i = \begin{cases} +\varphi(N-i) \text{ for } R_i = 1, \\ -\varphi(N-i) \text{ for } R_i = 0. \end{cases} \tag{8.4.9}$$

The number i should necessarily reflect the history of acquiring the data $\mathbf{L}_i \to R_i$, i.e. the numbers i ($i = 1, \ldots, N$) correspond to the discrete time moments t_i of the realization of the control objectives $\mathbf{Z}_i^*$ in the situation $\mathbf{X}_i$, and $t_i < t_{i+1}$.

Now total potential (8.4.7) is going to reflect better the 'fresh' data on the operation of the control system and, owing to the introduction of function (8.4.9), 'forget' the 'old' data. The intensity of this 'forgetting' should depend on the rate of the evolution of the object and the medium and reflect the latter. For instance, the function

$$\varphi(i) = \begin{cases} 1 \text{ for } 1 \leqslant i \leqslant N_1, \\ 0 \text{ for } N_1 < i \end{cases}$$

allows considering only the last N_1 members of series (8.4.3). The less N_1, the more intense the forgetting. And conversely, at $N_1 = N$ all available data (8.3.3) are taken into consideration equally and we get case (8.4.5).

Comments on References to Chapter 8

The correction of the control system is a form of its adjustment to the changing factors of the medium. This adjustment may be both of parametrical and of structural nature. In the former case the parameters of the control system are changed (e.g. the parameters of the object's model, see § 8.1), and in the latter case its structure is changed (see, for instance, § 8.2).

The mathematical apparatus used for the parameter correction in § 8.1 is in fact the apparatus of adaptive identification described with the relevant references in Chapter 4.

The structural adjustment considered in § 8.2, 8.3 is of alternative nature, i.e. it makes it possible to determine at each step one of the beforehand specified alternatives (the models in § 8.2 and the objects in § 8.3). The mathematical apparatus here is filtering (smoothing) the random processes. One may introduce oneself into this apparatus by any textbook on random processes, e.g. [1-4].

The presented in § 8.4 method to produce the solving function is called the method of potential functions [5]. However, generally speaking, other methods of pattern recognition may be applied for these purposes as well. The reader is recommended for the initial acquaintance with the pattern recognition problem to take one of the books [6, 7]. The textbooks [5, 8-10] deal with various methods of recognition in more detail.

References

1. Sveshnikov, A.A. *Applied Methods of the Theory of Random Functions.* Nauka, Moscow, 1968 (in Russian).
2. Pugachev, V.S. *The Theory of Random Functions.* Fizmatgiz, Moscow, 1962 (in Russian).
3. Bendat, J.S. *Principles and Applications of Random Noise Theory.* Wiley, Inc., New York, 1960.
4. Tikhonov, V.I. *Statistical Radio Engineering.* Sov. radio, Moscow, 1966 (in Russian).
5. Aizerman, M.A., Braverman, E.M., Rozonoer, L.I. *The Method of Potential Functions in the Theory of Trained Machines.* Nauka, Moscow, 1970 (in Russian).
6. Arkad'ev, A.G., Braverman, E.M. *Training the Machine to Classify Objects.* Nauka, Moscow, 1971 (in Russian).
7. Bongard, M.M. *The Problem of Recognition.* Nauka, Moscow, 1967 (in Russian).
8. Vapnik, V.N., Chervonenkis, A.Ya. *The Theory of Pattern Recognition: the Statistical Problems of Training.* Nauka, Moscow, 1974 (in Russian).
9. Ivakhnenko, A.G. (ed.). *Perceptron—the System for Pattern Recognition.* Naukova dumka, Kiev, 1975 (in Russian).
10. Zagoruiko, N.G. *The Methods of Recognition and Their Applications.* Sov. radio, Moscow, 1972 (in Russian).

Supplement. The Optimum Seeking Methods

The problem of optimization may be the most common computational problem of our time. As a matter of fact, practically all problems of control may be reduced to seeking the optimum for a function or functional, and each stage of the complex object control makes it necessary to solve optimization problems of several modifications. Therefore, it is very important to wield the methods to solve them. There are many such methods. The most efficient of them are described below; they were not once used for the problems of control.

S.1. Formulation of the Problem

The problem to optimize the function $Q = Q(\mathbf{X})$ of many variables $\mathbf{X} = (x_1, \ldots, x_k)$ is formulated as follows:

$$Q(\mathbf{X}) \to \min_{\mathbf{X} \in \Omega} \Rightarrow \mathbf{X}^*, \tag{S.1.1}$$

where Ω is the set of admissible values of the variable $\mathbf{X}$, and $\mathbf{X}^*$ is the solution of the optimization problem, i.e.

$$Q(\mathbf{X}^*) = \min_{\mathbf{X} \in \Omega} Q(\mathbf{X}),$$

where $\mathbf{X}^* \in \Omega$. The set Ω is usually specified with a system of equalities and inequalities in the form

$$\Omega : \begin{cases} g_i(\mathbf{X}) = 0, & i = 1, \ldots, p < k, \\ h_j(\mathbf{X}) \geqslant 0, & j = 1, \ldots, q, \end{cases}$$

where $g_i(\mathbf{X})$ and $h_j(\mathbf{X})$ are some given scalar functions. These functions, as well as the function $Q(\mathbf{X})$ may be specified not only analytically, but also algorithmically, i.e. by means of an algorithm permitting the calculation of the values of these functions for any given point $\mathbf{X}$.

The seeking method to solve problem (S.1.1) may be realized according to a recurrent expression showing the transfer from the $(N-1)$-th to the N-th approximation:

$$\mathbf{X}_N = \mathbf{X}_{N-1} + \Delta\mathbf{X}_N.$$

The step $\Delta\mathbf{X}_N$ depends on the situation in the point $\mathbf{X}_{N-1}$; $\Delta\mathbf{X}_N = \varphi(\mathbf{X}_{N-1})$, where φ is the seeking algorithm defining the recurrent optimization process. The algorithm is almost completely deter-

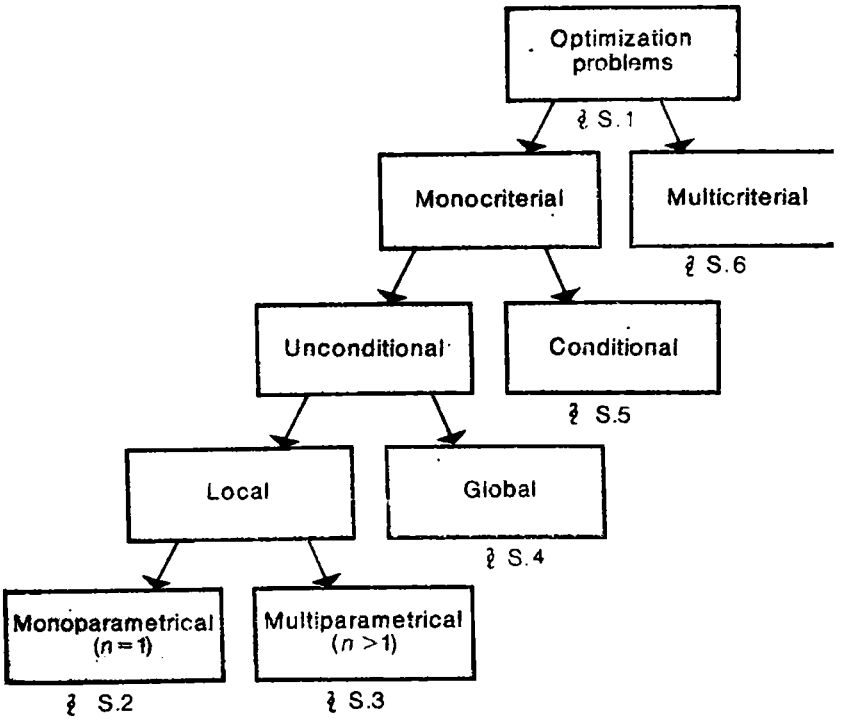

Fig. S.1.1. Structure of optimum seeking problems

mined by the characteristics of the optimization problem. Let us consider the main of them.

1. Depending on the number of minimized functions, i.e on the dimension of the function Q (X), there are monocriterial (see § S.2-S.5) and multicriterial vector (see § S.6) problems of optimization.

2. Depending on whether or not the solution X* of the problem belongs to the boundary $\overline{\Omega}$ of the area Ω, there are unconditional ($X^* \notin \overline{\Omega}$, presented in § S.2-S.4) and conditional ($X^* \in \overline{\Omega}$, considered in § S.5) optimization problems.

3. Depending on the number of local extrema, there are local, or monoextremal (considered in § S.2-S.3), and global, or multiextremal (discussed in § S.4), problems.

4. Depending on the number n of the optimized parameters, there are monoparametrical ($n = 1$, see § S.2) and multiparametrical ($n > 1$, see § S.3) optimization problems.

The cited classification of optimum seeking problems is shown diagrammatically in Fig. S.1.1 indicating the numbers of the sections where the main methods to solve the specified problems are presented.

S.2. Monoparametrical Local Search ($n = 1$)

The problem of monoparametrical local search may be put down in the form

$$Q(x) \to \min_x \Rightarrow x^*. \qquad \text{(S.2.1)}$$

First, let us consider the case when there are no disturbances. If the function $Q(x)$ is differentiable, then it is sufficient for the solution of problem (S.2.1) to solve another problem

$$\frac{dQ(x)}{dx} = f(x) = 0. \qquad \text{(S.2.2)}$$

But if the function is non-differentiable or its computation is very involved (e.g. in case the function is specified algorithmically), it is possible to determine $f(x)$ by means of the point estimate

$$\hat{f}(x) = \frac{\widehat{dQ}(x)}{dx} = \frac{1}{2g}[Q(x+g) - Q(x-g)], \qquad \text{(S.2.3)}$$

where $g > 0$ is a small enough value (the basis of the estimate).

As it is seen from (S.2.2), the optimization problem may be reduced to the problem to find the root of the equation $f(x) = 0$. Therefore, it is possible for its solution to make use of the numerous recurrent methods elaborated for the calculation of such roots. Let us note that since the function $Q(x)$ has only one extremum, the number of the roots of equation (S.2.2) does not exceed unity.

1. The *method of tangents* is applied for the convex function $Q(x)$, which corresponds to the monotonicity of its derivative $f(x)$. The algorithm of the method at each step is reduced to the linear presentation of $f(x)$ in each point x_i and to the solution of the equation

$$f(x) \approx f(x_i) + (x - x_i) f'(x_i) = 0,$$

whose root is the next approximation

$$x_{i+1} = x_i - f(x_i)/f'(x_i).$$

Obviously, it is necessary for that to determine the second derivative $f'(x) = Q''(x)$ of the minimized function. Its point estimate is

$$\hat{f}'(x) = \frac{1}{g}\left[f\left(x + \frac{g}{2}\right) - f\left(x - \frac{g}{2}\right)\right]$$

$$\approx \frac{1}{g^2}[Q(x+g) - 2Q(x) + Q(x-g)],$$

where the basis of the estimate equals $g/2$.

Consequently, the algorithm of tangents in solving problems of unidimensional optimization may be put down as follows:

$$x_{i+1}=x_i-\frac{g\,[Q(x_i+g)-Q(x_i-g)]}{2\,[Q(x_i+g)-2Q(x_i)+Q(x_i-g)]}.$$

Evidently, to make one optimization step according to the method of tangents, it is necessary to determine the minimized function in the three points x_i, $x_i \pm g$. Let us recall that this method is always convergent, i.e. $\lim\limits_{i\to\infty} x_i = x^*$, only for convex functions $Q(x)$.

2. The *method of chords* implies the approximation of the function $f(x)$ in the vicinity of the root by a chord connecting the points on both sides of the root. Suppose $x_i' < x_i''$ are two points in the vicinity of the root at the i-th step while (for the minimization case)

$$f(x_i') < 0, \quad f(x_i'') > 0, \tag{S.2.4}$$

i.e. they lie to both sides of the root x^*. This chord is described by the equation

$$f(x)\frac{1}{x''_i-x'_i}[(x-x'_i)f(x''_i)+(x''_i-x)f(x'_i)].$$

Making this expression equal to zero, we get x_{i+1}, the next approximation of the root. Let us form x'_{i+1} and x''_{i+1} for the next step in the following way:

$$\begin{aligned} &x'_{i+1}=x'_i \text{ and } x''_{i+1}=x_{i+1} \text{ if } f(x_{i+1})>0 \\ &x'_{i+1}=x_{i+1} \text{ and } x''_{i+1}=x''_i \text{ if } f(x_{i+1})<0. \end{aligned} \tag{S.2.5}$$

Here and in the preceding formula one should apply instead of the function $f(x)$ its point estimate according to (S.2.3). The method of chords converges for the non-convex functions $Q(x)$ as well.

3. The *method of dichotomy* allows finding the estimate for the root of $f(x)$ (and therefore the minimum for $Q(x)$) by means of dividing in half the segment within which the root is. Suppose the ends of such a segment, as for the method of chords, are $x_i' < x_i''$ meeting requirements (S.2.4). Then according to this algorithm

$$x_{i+1}=\frac{1}{2}(x'_i+x''_i).$$

A new segment $[x'_{i+1}, x''_{i+1}]$ is formed for the next step according to rule (S.2.5), and so on. There is no need for this method in calculating the value of the derivative $f(x)$, it is sufficient to determine its sign. The method converges for any monoextremal function $Q(x)$.

Let us note that in the realization of the presented methods it is necessary for accurate determination of the extremum to

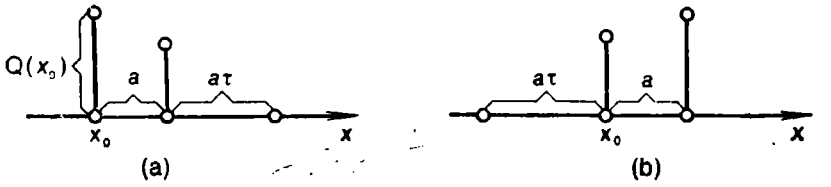

Fig. S.2.1. Location of search points at the initial stage of search (the vertical segments correspond to the function values)

lessen the basis of the estimate g with the advance of t. Indeed, the fact that the point estimate of the derivative (S.2.3) for the minimized function is equal to zero, i.e.

$$Q(x_t + g) - Q(x_t - g) = 0,$$

determines the extremum only if $g \to 0$.

Now let us consider the methods of monoparametrical optimization which make no use of the $f(x)$ derivative.

4. The *golden section method.* Let us consider the algorithm for this method starting from the initial point x_0. Let us make an arbitrary step a and find there the value of the minimized function $Q(x_0 + a)$. If $Q(x_0 + a) < Q(x_0)$, then the function is to be found next in the point $x_0 + a(1 + \tau)$, where $\tau = (1 + \sqrt{5})/2 \approx \approx 1.62$ (Fig. S.2.1*a*); otherwise it is to be found in the point $(x_0 - a\tau)$ (Fig. S.2.1*b*). Now we have three points on the axis x (let us denote them $x^1 < x^2 < x^3$) with the corresponding values of the minimized function Q_1, Q_2, and Q_3, i.e. $Q_j = Q(x^j)$ $(j = 1, 2, 3)$. Thus in the first case $x^1 = x_0$, $x^2 = x_0 + a$, $x^3 = x_0 + a(1 + \tau)$, in the second case $x^1 = x_0 - a\tau$, $x^2 = x_0$, $x^3 = x_0 + a$.

Two situations are possible here:

(1) The least value of the function belongs to the middle of these three points, i.e.

$$\min_{j=1,2,3} Q_j = Q_2. \qquad \text{(S.2.6)}$$

Then obviously the minimum of the function $Q(x)$ is between the points x^1 and x^3. The next value of the minimized function should be calculated for the point $x^4 = x^1 + x^3 - x^2$.

One extreme point out of the four x^j $(j = 1, \ldots, 4)$ is rejected so that the middle point of the rest would have the least value of the function $Q(x)$. The new three points are again denoted as $x^1 < x^2 < x^3$. Relationship (S.2.6) holds true for them. Therefore, the same rule is applied here again. Thus, the position of the extremum is located more accurately. The criterion for the end of the process may be, as usual, the accuracy of the estimate of the extremal point, i.e. the difference $(x^3 - x^1)$.

(2) In case $\min\limits_{j=1,\,2,\,3} Q_j \neq Q_2$, i.e. the least is either Q_1 or Q_3, add a point

$$x^4 = \begin{cases} x^1 - \tau(x^2 - x^1) \text{ if } \min\limits_{j=1,\,2,\,3} Q_j = Q_1, \\ x^3 + \tau(x^3 - x^2) \text{ if } \min\limits_{j=1,\,2,\,3} Q_j = Q_3. \end{cases}$$

Select three out of the four points $x^1, \ldots, x^4$ as follows. If the least value of the function belongs to one of the extreme points, reject the point with the greatest value (this is going to be the other extreme point), and denote the rest as $x^1 < x^2 < x^3$. This is the second situation, so further apply to (2). Otherwise reject one of the extreme points so that the middle of the rest three points would have the least value of the function. Denoting these points similarly, we reveal that this is the first situation, and apply (1).

Consequently, we come to such a value $x^* = x^2$ that is the solution of the problem, where x^2 is taken from the last three figures $x^1 < x^2 < x^3$. (Let us note that in principle we could take some other τ, but the indicated τ gives the greatest rate of convergence for the optimization process.)

The methods presented above may be applied only when there are no random disturbances superimposed in the quality function $Q(x)$. When there are such disturbances, we should take some other extremum seeking methods (their algorithms are often called adaptation algorithms thus emphasizing that the seeking algorithm adjusts, adapts, as it were, to the uncertainty related to the presence of random disturbances).

5. The *method of stochastic approximation* is applied to minimize the functions whose value is determined with a random disturbance. It may be put down in the form

$$x_{i+1} = x_i - \frac{a_{i+1}}{2g_{i+1}} [Q(x_i + g_{i+1}) - Q(x_i - g_{i+1})], \tag{S.2.7}$$

where the positive figures a_i, g_i should meet the following conditions:

$$\lim_{i\to\infty} a_i = \lim_{i\to\infty} g_i = 0, \quad \sum_{i=1}^{\infty} a_i = \infty, \quad \sum_{i=1}^{\infty} \left(\frac{a_i}{g_i}\right)^2 < \infty. \tag{S.2.8}$$

These conditions (they are called Dvoretski's conditions) are met, for instance, by the series $a_i = 1/i^\alpha$, $g_i = 1/i^\beta$ where the parameters α and β are constrained with $0 < \alpha < 1$, $\beta > 0$, $\alpha - \beta > 1/2$. The extremum seeking process for the function $Q(x)$ in that case converges regardless of the level of the random disturbances superimposed on that function.

6. The *adaptive method* is related to an adaptive change of the step Δx in the seeking process. The essence of the algorithm is that when the minimized function decreases, a step is made in the same direction, and when it increases, in the reverse direc-

tion:

$$\Delta x_{i+1} = \begin{cases} \gamma_1 \Delta x_i \text{ if } \Delta Q_i < 0, \\ -\gamma_2 \Delta x_i \text{ if } \Delta Q_i \geqslant 0, \end{cases} \tag{S.2.9}$$

where $\Delta x_i = x_i - x_{i-1}$; $\Delta Q_i = Q(x_i) - Q(x_{i-1})$, while $\gamma_1 \geqslant 1$, $0 < \gamma_2 < 1$ are the parameters characterizing the increase in the step (γ_1) while travelling in the 'successful' direction and the decrease (γ_2) in the reverse, i.e. when 'passing over' the extremum. The algorithm is convergent for $\gamma_1 \gamma_2 < 1$.

S.3. Multiparametrical Local Search ($n > 1$)

The descent algorithms are most common among the algorithms of local search. They differ in the procedure of descent along the selected direction. Let us consider them first. We can present a step $\Delta \mathbf{X}_i$ in the form

$$\Delta \mathbf{X}_i = a_i \mathbf{V}_i, \tag{S.3.1}$$

where $\mathbf{V}_i$ is the unit vector of the direction for the i-th step; a_i is the value of the step. Different methods of optimization depend on the techniques to produce the vector $\mathbf{V}_i$.

1. The *coordinate* (Gauss-Seidel) *method* is characterized by that the directions of $\mathbf{V}_i$ correspond to the directions of the coordinate axes, i.e. $\mathbf{V}_i = \mathbf{e}_j$ ($j = 1, \ldots, k$), where $\mathbf{e}_j$ is the j-th unit vector, i.e. the direction of the j-th coordinate. The next after j-th step is along the $(j+1)$-th coordinate, and so on up to the k-th coordinate, after which a step is made again along the first coordinate, i.e. along $\mathbf{e}_1$, etc. Let us note that here and further on the term 'along' may mean travel not only in the indicated direction, but in the opposite as well.

2. The *gradient method* (the method of the quickest descent) is characterized by that the direction $\mathbf{V}_i$ is antigradient, i.e.

$$\mathbf{V}_i = -\text{dir} \nabla_{\mathbf{X}} Q(\mathbf{X}_{i-1}),$$

where 'dir' denotes the direction of the vector: $\text{dir } \mathbf{Z} = \mathbf{Z}/|\mathbf{Z}|$, and $\nabla_{\mathbf{X}}$ is the gradient (del operator), i.e. a vector with the components equal to the partial derivatives of the function the symbol belongs to:

$$\nabla_{\mathbf{X}} Q = \left(\frac{\partial Q}{\partial x_1}, \ldots, \frac{\partial Q}{\partial x_k} \right).$$

The optimized parameters in this method are changed so that the minimized function first decreases in the quickest way. It is necessary for that, however, to know all the partial derivatives

or their estimates. The point estimates here have the form

$$\frac{\widehat{\partial Q}}{\partial x_j}=\frac{1}{g}[Q(\mathbf{X}+g\mathbf{e}_j)-Q(\mathbf{X})],$$
$$\frac{\partial Q}{\partial x_j}=\frac{1}{2g}[Q(\mathbf{X}+g\mathbf{e}_j)-Q(\mathbf{X}-g\mathbf{e}_j)]. \qquad \text{(S.3.2)}$$

where $g > 0$ is the basis of the estimate.

It is easy to note that for $g \to 0$ both estimates converge to the accurate value of the derivative (naturally, if it exists). The first estimate in (S.3.2) is biased, and the second unbiased, but the difficulty in determining the latter is twice as great since the function $Q(\mathbf{X})$ for it should be determined in two points $\mathbf{X} \pm g\mathbf{e}_i$.

3. The *method of local random search* is characterized by that the direction of the vector $\mathbf{V}_i$ is random: $\mathbf{V}_i = \Xi$, where Ξ is a unit random vector evenly distributed in all the directions of the space $\{\mathbf{X}\}$ of the parameters.

To realize any of the discussed optimum seeking methods, it is necessary to specify the value of the step a_i (see (S.3.1)). To determine a_i, one should make use of the following obvious suggestion. The value of $Q(\mathbf{X}_i)$ should be as small as possible, i.e. it is necessary to solve the problem

$$Q(\mathbf{X}_i)=Q(\mathbf{X}_{i-1}+a_i\mathbf{V}_i) \to \min_{a_i} \Rightarrow a_i^*, \qquad \text{(S.3.3)}$$

where a_i^* is the optimal value of the step. Let us note that, generally speaking, this may be a negative value as well (for instance, in the case of random search). This is a problem of the monoparametrical optimization that may be solved by the common methods of unidimensional search described above (see § S.2), and it is usually called for by descent.

4. The *method of multidimensional stochastic approximation*. The algorithm of this method may be put down in the form

$$\mathbf{X}_{i+1} = \mathbf{X}_i - a_{i+1}\,\widehat{\nabla}_{\mathbf{X}}\,Q(\mathbf{X}_i),$$

where $\widehat{\nabla}_{\mathbf{X}}\,Q(\mathbf{X}_i)$ is a point estimate of the gradient whose components are estimates of the partial derivatives similar to (S.3.2):

$$\frac{\widehat{\partial Q}}{\partial x_j}=\frac{1}{2g_{i+1}}[Q(\mathbf{X}_i+g_{i+1}\mathbf{e}_j)-Q(\mathbf{X}_i-g_{i+1}\mathbf{e}_j)],$$

and the variable coefficients a_i and g_i should satisfy the Dvoretski's conditions (S.2.8).

S.4. Algorithms for Global Search

In case the minimized function $Q(\mathbf{X})$ has more than one extremum and it is necessary to find the least of them, neither of the described above algorithms solves the problem. The algorithms of

the global search are used for that. Let us consider some of them.

1. The *method of the global random search*, applied correctly, is an efficient enough means for global extremum search. It consists in the following. Random points $X_1, X_2, \ldots$ are generated in the search area Ω in compliance with a specified, for instance, normal, distribution density $P(X \mid W, \sigma^2)$. Here W is the mathematical expectation of the distribution, and σ^2 is the maximal dispersion of its components. The essence of the algorithm consists in that W is situated in the point with the least found earlier value of the minimized function, i.e. $W_i = X^*_i$, for which $Q(X^*_i) = Q^*_i = \min_{j=1,\ldots,i} Q(X_j)$.

The algorithm for such search is put down in the recurrent form as

$$X^*_i = X^*_{i-1}, \quad Q^*_i = Q^*_{i-1} \text{ if } Q(X_i) \geqslant Q^*_{i-1},$$

$$X^*_i = X_i, \quad Q^*_i = Q(X_i) \text{ if } Q(X_i) < Q^*_{i-1}.$$

It is clear that the algorithm guarantees finding the global extremum if the distribution density of random points is not zero anywhere in the search area Ω, i.e.

$$P(X \mid W, \sigma^2) \neq 0 \quad \text{for} \quad X \in \Omega.$$

Indeed, since the probability to hit the vicinity of the global extremum is incidentally finite, this event is going to occur sooner or later. But rather later than sooner since the algorithm converges slowly, and therefore in the process of its realization it is necessary to take measures to increase the rate of convergence.

Adaptation of dispersion σ^2 is applied for the purpose in the process of search: the dispersion is to be decreased after unsuccessful random trials and increased after successful, i.e. $\sigma_i = \sigma_{i-1} + \Delta\sigma_i$, where the increment $\Delta\sigma_i = \varphi(\sigma_{i-1}, \Delta Q_{i-1})$ is determined by the adaptation algorithm. For instance,

$$\sigma_i = \begin{cases} \gamma_1 \sigma_{i-1} & \text{for } \Delta Q_{i-1} > 0, \\ \gamma_2 \sigma_{i-1} & \text{for } \Delta Q_{i-1} \leqslant 0. \end{cases}$$

(Here the coefficients $\gamma_1 < 1$ and $\gamma_2 \geqslant 1$ are selected for the concrete situation.) The correct choice of these coefficients considerably determines the success of the global search. If they are close to unity, the adaptation effect is inessential, while the search is long but successful, i.e. the global extremum is going to be found. But if the coefficients differ from unity appreciably, then the process will end soon, but it may happen that the global extremum will not be found. The less γ_j, the sooner the search ends, and the greater the risk to 'miss' the global extremum. Selecting γ_1 and γ_2, one should reasonably compromise between the period of search and its reliability.

2. The *method of the smoothing global search* is effectively applied when the minimized function is multiextremal due to the influence of some oscillating (not necessarily periodical) component $Q_2(X)$ superimposed on a 'good' monoextremal function

Q_1 (X):

$$Q\ (\mathbf{X}) = Q_1\ (\mathbf{X}) + Q_2\ (\mathbf{X}).$$

Such separation is unknown a priori. The problem is to 'smooth' Q_2 (X) and by the virtue of this reduce the multiextremal problem to the local minimum seeking problem for the monoextremal function Q_1 (X), considered in § S.2-S.3.

The procedure to smooth the initial function Q (X) should be of averaging nature. The averaging should take into account a weighting function ρ (Y), where $\mathbf{Y} \in \Omega$ is the vector deviation from X. Therefore it should be:

—non-negative, i.e. $\rho\ (\mathbf{Y}) \geqslant 0$ for $\mathbf{Y} \in \Omega$;

—normalized, i.e. $\int\limits_{\Omega} \rho\ (\mathbf{Y})\ d\mathbf{Y} = 1$;

—δ-like by one of its internal parameters q, i.e. $\lim\limits_{q \to 0} \rho(\mathbf{Y}) =$ $= \delta\ (\mathbf{Y})$, where $\delta\ (\cdot)$ is a delta-function.

The Gaussian function

$$\rho\ (\mathbf{Y}) = \frac{1}{(2\pi)^{n/2}\ q^n} \exp\left\{-\frac{1}{2q^2} \sum_{i=1}^{k} y^2{}_i\right\}$$

and the multidimensional 'witch of Agnesi'

$$\rho\ (\mathbf{Y}) = l/(1 + q\ |\ \mathbf{Y}\ |^2),$$

where l is the normalizing factor, are such functions.

Now the procedure to average the minimized function may be presented as the k-fold integration with the weight ρ:

$$\overline{Q}\ (\mathbf{X}) = \int\limits_{\Omega} Q\ (\mathbf{X}+\mathbf{Y})\ \rho\ (\mathbf{Y})\ d\mathbf{Y}, \qquad \text{(S.4.1)}$$

where $\overline{Q}$ (X) is the value of the function Q (X) averaged over Ω.

It is very difficult to take this integral correctly. Therefore it should be estimated by the Monte Carlo method. The weighting function ρ (Y) may be considered as the distribution density for the random vector Y. Then (S.4.1) may be put down as

$$\overline{Q}\ (\mathbf{X}) = M_{\mathbf{Y}} Q\ (\mathbf{X} + \mathbf{Y}),$$

where the mathematical expectation $M_{\mathbf{Y}}$ is meant for the random vector Y with the distribution density ρ (Y).

Suppose $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ are concrete realizations of this random vector. Then we can get the Monte Carlo estimate for integral (S.4.1)

$$\hat{\overline{Q}}_N(\mathbf{X}) = \frac{1}{N} \sum_{i=1}^{N} Q\ (\mathbf{X}+\mathbf{Y}_i). \qquad \text{(S.4.2)}$$

Obviously, this estimate converges to the correct value of the integral:

$$\lim_{N\to\infty} \hat{\overline{Q}}_N(\mathbf{X}) = \overline{Q}(\mathbf{X}).$$

Since the averaged function $\overline{Q}(\mathbf{X})$ is monoextremal (because $\overline{Q}(\mathbf{X}) \approx Q_1(\mathbf{X})$), any local search algorithms may be applied to it, expression (S.4.2) giving an estimate of the function in the point X. Having this estimate, we can use any of the above considered local search algorithms, i.e. solve the multiextremal problem by a local method. However, it should be borne in mind that each realization of the estimate for the function smoothed in this way is a random value only the average of which is equal to $\overline{Q}(\mathbf{X})$. Taking the estimates $\hat{\overline{Q}}$ in the seeking process is equivalent to search in the presence of random disturbances superimposed on the function $\overline{Q}(\mathbf{X})$. The effect of these disturbances is in inverse proportion to the volume N of the sample (see (S.4.2)), which allows managing the disturbance or, rather, decreasing its dispersion by increasing N.

The parameter q determines the degree of weighting function 'diffusion'. The greater q, the greater the averaging and the 'rougher' the solution to the problem. The averaging at small q is low. Evidently, this parameter should be changed in the process of search. It should be large at the beginning and tend to zero towards the end of the search. With a function $g(t)$ selected properly the process of global extremum seeking converges rapidly enough.

S.5. Conditional Optimization

The difficulty of the problem of conditional optimization lies in that its solution $\mathbf{X}^*$ is on the boundary $\overline{\Omega}$ of the admissible area Ω, i.e. $\mathbf{X}^* \in \overline{\Omega}$. The difficulty is negotiated in two ways. On the one hand, it is possible to make use of the specificity of the area Ω, i.e. take into account the kind of functions determining it. This is what is done by the methods of mathematical programming. Some of them are considered in § 6.4. However, another approach is possible (and it is more expedient in a number of cases) which consists in reducing the conditional optimization problem to unconditional. The seeking algorithms are the same in this case (see § S.2-S.4), and the main attention is focused on the procedure of transforming the problem into an unconditional one.

Let us consider the main approaches to that.

1. The *substitution method.* It is used when there are constraints of the equality type

$$\Omega : g_i(\mathbf{X}) = 0, \quad i = 1, \ldots, p < k. \qquad \text{(S.5.1)}$$

These constraints are a system of p equations in k unknowns. Let us solve this system with regard to p variables through the

rest $(k - p)$ of them and substitute these expressions into Q (X). The dimension of the solved problem is going to decrease by p and it becomes unconditional. This is the way they do in case it is easy to be performed, for instance, in linear programming problems of the (6.2.1) kind thus transformed into form (6.4.1).

Obviously, such a transformation is realizable only if the system of equations (S.5.1) may be solved analytically. As we can recall, the equations may be transcendent in the general case, so the possibilities of such an approach are limited.

However, the recurrent method to solve the optimization problems helps negotiate this obstacle in the following way. Suppose we split the vector of the sought parameters X into two:

$$\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2) \qquad \text{(S.5.2)}$$

where

$$\mathbf{X}^1 = (x_1, \ldots, x_{k-p}); \quad \mathbf{X}^2 = (x_{k-p+1}, \ldots, x_k). \qquad \text{(S.5.3)}$$

Let us minimize the quality function by means of $\mathbf{X}^1$ while meeting conditions (S.5.1) by the proper selection of $\mathbf{X}^2$. The latter is easy: it is sufficient to solve the system of p equations

$$g_i(\mathbf{X}^1, \mathbf{X}^2) = 0, \quad i = 1, \ldots, p \qquad \text{(S.5.4)}$$

with regard to p unknowns $\mathbf{X}^2$ at the given values $\mathbf{X}^1$. It may be done with any numerical method (see the comments on § S.5). Therefore, for any specified $\mathbf{X}^1$ it is always possible to determine such $\mathbf{X}^2$ for which conditions (S.5.1) for Ω will be met.

Now the initial problem of conditional optimization is transformed into a problem of the unconditional optimization

$$Q(\mathbf{X}^1, \mathbf{X}^2) \to \min_{\mathbf{X}^1},$$

where the vector $\mathbf{X}^2$ is to be determined at each step of unconditional optimization as the numerical solution of system (S.5.4) for the given values $\mathbf{X}^1$.

Let us apply this approach for the case when there are constraints of the inequality type

$$\Omega : h_j(\mathbf{X}) \geqslant 0, \quad j = 1, \ldots, q.$$

The system of equations to be solved in this case is formed from the equations of violated inequalities

$$h_j(\mathbf{X}^1, \mathbf{X}^2) = 0, \quad j = z_1, \ldots, z_\tau, \qquad \text{(S.5.5)}$$

where $z_1, \ldots, z_\tau$ are the numbers of the violated constraints, i.e. inequalities.

Here the dimension of the determined vector $\mathbf{X}^2$ is equal to τ, the number of the violated constraints. The number τ changes in the process of solving the system (S.5.5). The dimension of $\mathbf{X}^2$ changes likewise. The process ends when $\tau = 0$.

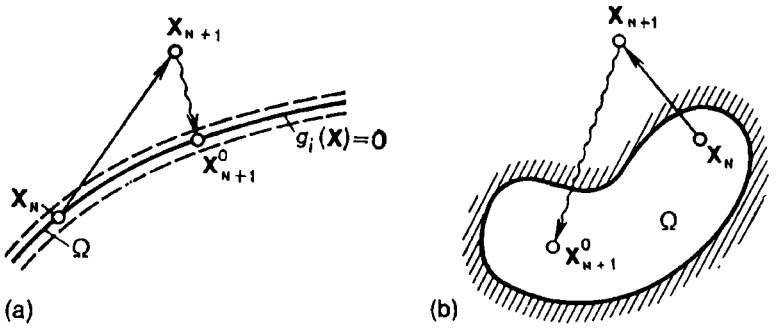

Fig. S.5.1. Demonstration of operation of the projection method with constraints of the type of equality (*a*) and inequality (*b*)

The approach may be generalized rather easily for the case when there are constraints of two types:

$$\Omega : \begin{cases} g_i(X)=0, & i=1, \ldots, p. \\ h_j(X_j) \geqslant 0, & j=1, \ldots, q. \end{cases}$$

2. The *projection method.* The essence of the method may be reduced to the following. In case the process of solving problem (6.1.8) produces an intermediate value $X_{N+1} \notin \Omega$, i.e. violating the constraints Ω, it should be projected on Ω to meet the constraints, and carry on solving the problem from the projected point

$$X^0{}_{N+1} = \mathrm{pr}_\Omega X_{N+1} \in \Omega, \tag{S.5.6}$$

where the symbol pr_Ω means 'projection on Ω'. The procedure of projecting is demonstrated in Fig. S.5.1. If $\Omega : g_i(X) = 0$, then the projecting procedure is a descent on the constraint $g_i(X) = 0$ (see Fig. S.5.1*a*). If the area Ω is determined by inequalities (see Fig. S.5.1*b*), the projection gives one of the points $X^o \in \Omega$ of this area.

Let us first consider the case when $\Omega : g_i(X) = 0$ ($i = 1, \ldots, p < k$). The algorithm projecting from the point X on Ω consists in forming a process of moving to end in a small enough ε-vicinity of the area Ω, i.e. at

$$| g_i(X^o) | \leqslant \varepsilon_i, \quad i = 1, \ldots, p \tag{S.5.7}$$

(see the dash lines in Fig. S.5.1*a*), where ε_i is a given accuracy to meet the *i*-th constraint. The process may be organized by any of the local search methods described in § S.2-S.3. It is sufficient for this to minimize the quality function of the form

$$Q^o(X) = \sum_{i=1}^{p} | g_i(X) | \to \min_X \tag{S.5.8}$$

at the initial conditions $\mathbf{X} = \mathbf{X}_{N+1}$ (Fig. S.5.1*a*) with the end by condition (S.5.7). It is possible to put another problem:

$$Q^0(\mathbf{X}) = \max_{i=1, \ldots, p} | g_i(\mathbf{X}) | \to \min_{\mathbf{X}} \qquad \text{(S.5.8a)}$$

with the same criterion for stopping.

We can act similarly when the constraints of the inequality type are violated (see Fig. S.5.1*b*). It is necessary then to minimize the quality function of the type

$$Q^0(\mathbf{X}) = -\sum_{j \in \{v\}} h_j(\mathbf{X}) \to \min_{\mathbf{X}}, \qquad \text{(S.5.9)}$$

where $\{v\}$ is the set of the numbers of the violated constraints for which $h_j(\mathbf{X}) < 0$. The stopping criterion: $Q^0(\mathbf{X}) \leqslant \varepsilon > 0$. Similarly

$$Q^0(\mathbf{X}) = -\min_{j=1, \ldots, k_2} h_j(\mathbf{X}) \to \min_{\mathbf{X}} \qquad \text{(S.5.9a)}$$

with the same stopping criterion. Clearly, it is convenient to find the projection $\mathbf{X}^0$ of the point $\mathbf{X}$ by means of minimizing the degree of violation of the constraints, which may be realized by seeking optimization of some specially organized functions.

Consequently, when using the projection method to solve a conditional problem we solve two unconditional problems in succession, step by step, and they may be presented as a single problem

$$Q_*(\mathbf{X}) \to \min_{\mathbf{X}},$$

where

$$Q_*(\mathbf{X}) = \begin{cases} Q(\mathbf{X}) & \text{for } \mathbf{X} \in \Omega, \\ Q^0(\mathbf{X}) & \text{for } \mathbf{X} \notin \Omega, \end{cases} \qquad \text{(S.5.10)}$$

i.e. there should be a switch over from one minimized quality function to the other depending on meeting the conditions Ω.

3. The *method of fining functions*. This method is a direct extension of the projection method, but the projecting process is joined with minimizing the initial function. The idea of the method is to construct such a quality function whose minimum would be the solution of the overall problem. This function in the general case has the form

$$\tilde{Q}(\mathbf{X}) = Q(\mathbf{X}) + f_{\mathrm{A}}(g_1(\mathbf{X}), \ldots, g_p(\mathbf{X})) + f_{\mathrm{B}}(h_1(\mathbf{X}), \ldots, h_q(\mathbf{X})), \qquad \text{(S.5.11)}$$

where $f_{\mathrm{A}}(\cdot)$ and $f_{\mathrm{B}}(\cdot)$ are non-negative functions of the fines for violating the constraints of the equality and inequality types correspondingly constructed in the following way:

$$f_{\mathrm{C}} = \begin{cases} =0 & \text{for } \mathbf{X} \in \Omega, \\ >0 & \text{for } \mathbf{X} \notin \Omega, \end{cases} \quad (\mathrm{C} = \mathrm{A}, \mathrm{B}), \qquad \text{(S.5.12)}$$

i.e. there is no fine when the conditions Ω are met, and a positive fine for their violation is introduced.

The structure of the functions f_A and f_B may be very diverse. For instance, a fine for violating the constraints of the equality type may be put down as

$$f_A = \sum_{i=1}^{p} a_i \varphi (g_i (X)), \qquad (S.5.13)$$

where $\varphi (\cdot)$ is an even function with the following properties: $\varphi (x) \geqslant 0$, $\varphi (0) = 0$, $\varphi (x)$ for $x > 0$ is a monotonically increasing function. These are the characteristics, for instance, of the function $\varphi (x) = | x |^{\nu}$, where $\nu > 0$. A particular case of such fining function with $a_i = 1$ and $\nu = 1$ is presented above (S.5.8).

The function

$$f_A = \max_{i=1, \ldots, p} a_i \varphi (g_i (X))$$

generalizes function (S.5.8a).

Clearly, the difference between the methods of projections and of fining functions lies in that the latter method of accounting the constraints minimizes both the fine and the basic function simultaneously, while the former method takes care of the two procedures individually and independently from each other.

The coefficients a_i of fining function (S.5.13) are commonly called the fining coefficients. When $a \to \infty$, where

$$a = \min_{i=1, \ldots, p} a_i,$$

the solution of the unconditional problem

$$Q (X) + f_A (X) \to \min_X \Rightarrow X^*_A$$

coincides with the solution of the initial conditional problem

$$Q (X) \to \min_{X \in \Omega} \Rightarrow X^*,$$

where $\Omega : g_i (X) = 0 \quad (i = 1, \ldots, p)$. This obviously follows from the fact that if a grows, the fining function will be first to be minimized, i.e. the condition $X \in \Omega$ will be met, and only after that the function $Q (X)$ will be minimized. Therefore,

$$\lim_{a \to \infty} X^*_A = X^*.$$

Now let us consider the constraints of the inequality type. The fining functions in this case may have, for instance, the following form:

$$f_B (X) = \sum_{j=1}^{q} b_j \psi (h_j (X)), \qquad (S.5.14)$$

where $\psi (\cdot)$ is the function

$$\psi (z) = -\min (0, z), \qquad (S.5.15)$$

while

$$\min(0,\ z)=\begin{cases}0 & \text{for } z\geqslant 0.\\ z & \text{for } z<0.\end{cases}$$

Expression (S.5.15) is a positive function of fine for violating the constraint $z>0$. And the greater the violation of the constraint z, the greater the value of fine. The coefficients b_j play the part of the fining coefficients.

Another form of the fining function

$$f_{\mathbf{B}}(\mathbf{X})=\max_{j=1,\ \ldots,\ q} b_j\psi(h_j(\mathbf{X}))$$

is similar to (S.5.9a). When $b\to\infty$, where

$$b=\max_{i=1,\ \ldots,\ q} b_j,$$

the solution of the unconditional problem

$$Q(\mathbf{X})+f_{\mathbf{B}}(\mathbf{X})\to\min_{\mathbf{X}}\Rightarrow\mathbf{X}^*_{\mathbf{B}}$$

coincides with the solution of the conditional one, where

$$\Omega: h_j(\mathbf{X})\geqslant 0\quad (j=1,\ \ldots,\ q),\ \text{i.e.}$$

$$\lim_{b\to\infty}\mathbf{X}^*_{\mathbf{B}}=\mathbf{X}^*.$$

Now joining both kinds of the fining functions we get an unconditional problem of the form

$$\widetilde{Q}(\mathbf{X})=Q(\mathbf{X})+\sum_{i=1}^{p}a_i\varphi(g_i(\mathbf{X}))+\sum_{j=1}^{q}b_j\psi(h_j(\mathbf{X}))\to\min_{\mathbf{X}}.\quad\text{(S.5.16)}$$

Suppose the minimum of this function is located in the point $\mathbf{X}^*_{AB}$. In the general case $\mathbf{X}^*_{AB}\neq\mathbf{X}^*$. But with $a,\ b\to\infty$, where a and b were defined above, the solution of unconditional problem (S.5.16) coincides with the solution of the initial problem, i.e.

$$\lim_{a,\ b\to\infty}\mathbf{X}^*_{AB}=\mathbf{X}^*.\qquad\text{(S.5.17)}$$

This is what guarantees the success of the method of fining functions.

A few words on the organization of the solving process according to this method. In order to make use of property (S.5.17), it is necessary to start optimization at low values of the parameters a and b gradually increasing them in the seeking process up to their maximum values $a_{\max}$ and $b_{\max}$. The rate of growth is very important for these parameters. It should be optimal, i.e. not too great, but not too small.

Indeed, at a low rate of growth of the parameters a and b the transition to $a_{\max}$ and $b_{\max}$ is going to be too long. However at a very large rate of growth of a and b the function $\widetilde{Q}(\mathbf{X})$ becomes ravine-like (see Fig. 6.3.2*c*), which curbs its optimization and

makes the search longer. Best of all is to change the values a and b adaptively, i.e. increase them at a moment when there is an assurance that a solution is found close enough to X^*_{AB}.

S.6. Multicriterial Optimization

Until now we have been considering methods for solving the problem of optimizing a scalar function $Q(X)$. But practice shows that it is usually necessary to solve the problems where the criterion $Q(X)$ is not a scalar, but a vector

$$Q(X) = (q_1(X), \ldots, q_m(X)),$$

and the problem is to minimize simultaneously all the m criteria

$$q_i(X) \to \min_{X \in \Omega}, \quad i = 1, \ldots, m, \tag{S.6.1}$$

where Ω is the set of admissible solutions the same as for problem (S.1.1).

An example of a multicriterial problem may be the problem to synthesize a cocktail considered in § I.5 (before the moment of selecting one extremized criterion):

$$Q(U) = (-f_1(U), \quad -f_2(U), \quad f_3(U)) \to \min_{U \in \Omega},$$

where

$$\Omega: \begin{cases} \sum_{i=1}^{n} u_i = 1 \\ u_i \geqslant 0 \quad (i = 1, \ldots, n). \end{cases}$$

The multicriterial problem seems to be unsolvable at the first glance. Indeed, solving each of the m problems of scalar optimization (S.6.1) independently, we get m optimal local solutions

$$X^*_i : q_i(X^*_i) = \min_{X \in \Omega} q_i(X), \quad i = 1, \ldots, m.$$

Evidently, in the general case these solutions do not coincide, i.e. $X^*_i \neq X^*_j$ $(i \neq j)$, and therefore neither of the local solutions may be the solution of the initial multicriterial problem.

Let us introduce rather obvious notions of preference and equivalence. The state X_1 we shall call strictly more preferable than the state X_2 if $Q(X_1) \leqslant Q(X_2)$, i.e. among the equalities and inequalities $q_i(X_1) \leqslant q_i(X_2)$ $(i = 1, \ldots, m)$ there is always a strict inequality to be found. Let us denote this as $X_1 \succ X_2$ if $Q(X_1) \leqslant Q(X_2)$.

But if $Q(X_1) = Q(X_2)$, then we shall consider the states X_1 and X_2 to be equivalent: $X_1 \sim X_2$. Let us note that generally an equivalence does not correspond to an equality. It may prove in a particular case, when $m = k$, that the equivalence does correspond to the equality $X_1 = X_2$, but in the general case this is not so.

The symbols of preference ($\succ$) and equivalence ($\sim$) have the following two obvious properties:

(1) if $X_1 \succ X_2$ and $X_2 \succ X_3$, then $X_1 \succ X_3$—this is the transitive property of preference;

(2) if $X_1 \sim X_2$ and $X_2 \sim X_3$, then $X_1 \sim X_3$—this is the transitive property of equivalence.

Now let us introduce the notion of the 'criteria space' $\{Q\}$. This space has dimension m (according to the number of the criteria) and is formed by m orthogonal axis along which the values of $q_i = q_i(X)$ are measured. There is a tight interrelationship between the space $\{X\}$ of parameters and the space $\{Q\}$ of criteria. Each point of the space $\{X\}$ of the parameters is reflected into a certain point Q of the space of the criteria (the reverse does not always hold true).

The area Ω of the admissible states X in the space $\{X\}$ is therefore reflected into a certain area S of the space $\{Q\}$ of the criteria. It is clear from Fig. S.6.1 that any point of the cross-hatched set is more preferable than the point $Q(X_0)$ (let us recall that we consider a minimization problem). This allows us to select in the whole set S a subset S^* of the points $Q^* \in S^*$ for which no more preferable ones may be found (see the solid line in Fig. S.6.1). But it does not mean that the set S^* contains only the best points. Generally, among $Q \notin S^*$ there may be, and even certainly are the points in a sense better than some of those from S^*, but on the other hand among S^* there always exist more preferable points than $Q \notin S^*$.

The set $\Omega^* \in \{X\}$ that corresponds to S^* is usually called the Pareto set, negotiation set or compromise area. The significance of the set lies in that namely this set contains the solution of multicriterial problem (S.6.1).

Example. Suppose $k = m = 2$,

$$q_1 = x_1^2 + x_2^2, \quad q_2 = (x_1 - 1)^2 + (x_2 - 1)^2, \qquad (*)$$

$$\Omega : -1 \leqslant x_i \leqslant 1 \quad (i = 1, 2).$$

The area Ω and the Pareto set Ω^*—the segment of the straight line between the points of the minimum for both functions (*)—are shown in Fig. S.6.2*a*. The area S for this example is shown in Fig. S.6.2*b*, where the solid line corresponds to the set S^*. However, this 'solution' can never be satisfactory. First of all at least because there is a set of solutions while only one is needed. It seems at the first glance that this expands the possibilities of optimization. Indeed, it is always possible to introduce one more additional criterion $q_{m+1}(X)$ and solving the optimization problem on the Pareto set

$$q_{m+1}(X) \to \min_{X \in \Omega^*},$$

find an optimum X^* satisfying the $(m + 1)$-th criterion. But this is a wrong conclusion, since the optimum is obtained at the price of considerable deviation of the initial criteria from the optimum. This is well seen in Fig. S.6.1 where A is the optimum with respect

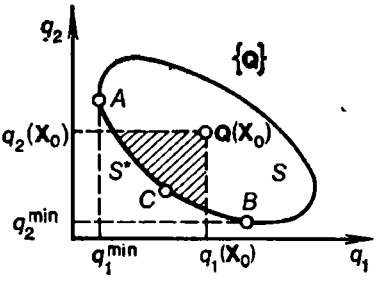

Fig. S.6.1. To determination of the Pareto set

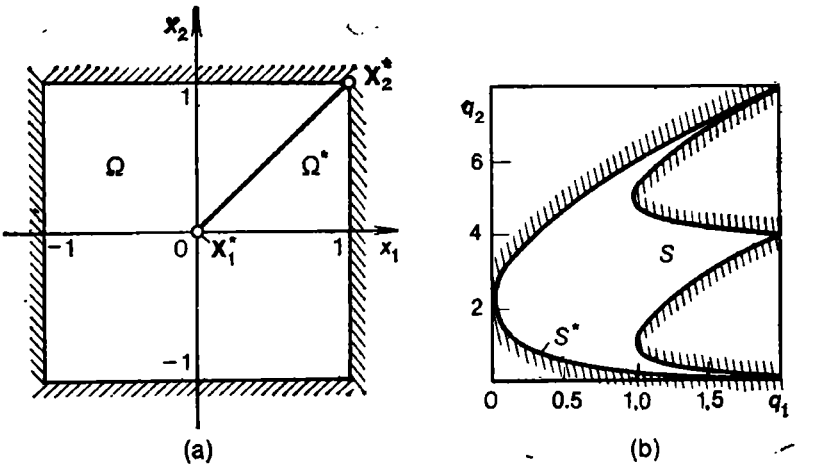

Fig. S.6.2. Example

to the first (and only first) criterion. But if we introduce a second criterion q_2, then the optimal point (suppose it is C) will be well far from the optima A and B, where B is the optimum for the criterion q_2. Therefore, the 'freedom' presented by the Pareto set has been 'bought' at the price of the essential deviation from the local optima.

To solve a multicriterial optimization problem unambiguously, it is *always* necessary to introduce some additional information I. Nothing can be said without it about the solution $\mathbf{X}^*$ except that it belongs to the Pareto set Ω^*. And only that. Therefore basically the process of solving the problem implies the search for and making use of the additional information I helping to reduce the multicriterial problem to a monocriterial one:

$$q(\mathbf{X}) = W(q_1(\mathbf{X}), \ldots, q_m(\mathbf{X}), I), \qquad \text{(S.6.2)}$$

where $q(\mathbf{X})$ is a scalar function whose minimum in Ω^* is the solution of the initial multicriterial problem; W is the convolution

of the function $q_i(\mathbf{X})$ $(i = 1, \ldots, m)$ using the needed information I.

The simplest method to reduce a multicriterial problem to a monocriterial one is its presentation in the form of the standard problem of the monocriterial mathematical programming

$$q_l(\mathbf{X}) \to \min_{\mathbf{X} \in L},$$

$$L: \begin{cases} \mathbf{X} \in \Omega, \\ q_i(\mathbf{X}) \leqslant \tilde{q}_i, \quad l \neq i = 1, \ldots, m, \end{cases}$$

where the information I is the assumption that it is necessary to minimize the l-th criterion while the rest $(m - 1)$ ones are to be bounded from above by the numbers $\tilde{q}_i$ $(i = 1, \ldots, m)$. This is the way the multicriterial problems were usually solved in the recent past. And this is how the cocktail problem was solved in § I.5. But the drawback of this method is obvious, since it is difficult to determine the 'main' criterion $q_l(\mathbf{X})$ and the numbers $\tilde{q}_i$ of the constraints. This makes us apply other methods.

Consequently, various methods to solve the multicriterial problem differ in the form and techniques to obtain the information I, as well as in the algorithm W of 'convolution' of the initial criteria into one global criterion $q(\mathbf{X})$.

All the available methods to solve the problem of the multicriterial optimization fall into three classes: the a priori, a posteriori and adaptive methods. Let us consider them.

The *a priori* method implies producing criterion (S.6.2) on the basis of only a priori data on the object. For instance,

$$q(\mathbf{X}) = \sum_{i=1}^{m} \lambda_i q_i(\mathbf{X}); \quad q(\mathbf{X}) = \prod_{i=1}^{m} q_i^{\lambda_i}(\mathbf{X})$$

etc., where the information I determines the values of the coefficients $\lambda_1, \ldots, \lambda_m$.

The 'convolution' of criteria may be the result of some other procedure:

$$q(\mathbf{X}) = \rho[Q(\mathbf{X}) - Q^*],$$

where $\rho(\cdot)$ is a certain weight function of the distance ($\rho(0) = 0$, $\rho(\cdot) \geqslant 0$); Q^* is the vector of minimal values of the criteria:

$$Q^* = (q_1^*, \ldots, q_m^*), \tag{S.6.3}$$

where $q_i^* = \min_{\mathbf{X} \in \Omega} q_i(\mathbf{X})$ $(i = 1, \ldots, m)$.

Here the solution is unambiguously determined by specifying the weighting function $\rho(\cdot)$. For instance,

$$q(\mathbf{X}) = \sum_{i=1}^{m} (q_i(\mathbf{X}) - q^*_i)^2.$$

It is possible to determine the solution $\mathbf{X}^*$ of the multicriterial problem proceeding from the condition of coincidence of a relative

'damage' for each of the criteria:

$$\frac{q_i(\mathbf{X})-q^*_i}{q^*_i}=\frac{q_j(\mathbf{X})-q^*_j}{q^*_j}, \quad i, j=1, \ldots, m.$$

The a priori method should be used when the volume of the a priori information is great enough to select unambiguously the technique of 'convolution' of all the m criteria into one (S.6.2). If not, then the a posteriori method is applied.

The *a posteriori* method implies the organization of collecting the information I before solving the problem. Let us first define the properties of the convolution function $W(\cdot)$ (S.6.2):

(1) The function $W(\cdot)$ should be invariant with regard to the shift transformation, i.e.

$$W(q_1, \ldots, q_m) = W(q_1 + c_1, \ldots, q_m + c_m),$$

where $c_1, \ldots, c_m$ are any figures. Indeed, the addition of a constant to any criterion should not affect the solution of a multicriterial problem. (This property is called additive.)

(2) The function $W(\cdot)$ should be invariant to any change of scale of any criterion, i.e.

$$W(q_1, \ldots, q_m) = W(d_1q_1, \ldots, d_mq_m),$$

where d_i are any positive numbers. (This property is called commutative.)

(3) The function $W(\cdot)$ should be symmetrical with regard to all the criteria, i.e. it should have the same structure with respect to $i = 1, \ldots, m$.

All these requirements are met by the function

$$W(\mathbf{X})=\sum_{i=1}\alpha_i\frac{q_i(\mathbf{X})-q^*_i}{q^{**}_i-q^*_i}, \qquad \text{(S.6.4)}$$

where $q^{**}_i = \max_{\mathbf{X}\in\Omega} q_i(\mathbf{X})$, while q^*_i were determined above (see (S.6.3)). It is easy to prove that by simple substitution. Each summand in (S.6.4) is none other than a normalized criterion which may change from 0 to 1.

Let us consider now the means to determine the weights α_i ($i = 1, \ldots, m$). First of all it is only natural to limit the freedom of selecting these parameters by the obvious conditions

$$\sum_{i=1}^{m}\alpha_i=1, \quad \alpha_i>0, \quad i=1, \ldots, m. \qquad \text{(S.6.5)}$$

To determine the parameters α_i, it is necessary to obtain the information I, e.g. by the experts' opinion poll as follows. The experts are given, to evaluate the preferences, several pairs of situations $\mathbf{X}_i^{(1)}$ and $\mathbf{X}_i^{(2)}$ ($i = 1, \ldots, p$) with the values of all criteria in these points, i.e. $q_j(X_i^{(1)})$ and $q_j(\mathbf{X}_i^{(2)})$ ($j = 1, \ldots, m$; $i = 1, \ldots, p$). The experts are to attribute to these pairs one of the three symbols $>$, $<$ or $\sim$. That is precisely the needed

information I producing in each case one of the three relationships

$$W(\mathbf{X}^{(1)}_i \lesseqgtr W(\mathbf{X}^{(2)}_i) \quad (i=1, \ldots, p).$$

Together with (S.6.5), the result is a system of linear equalities and inequalities with regard to the unknown weights α_i; solving the system by the standard methods, it is not difficult to find the area of the admissible values α_i ($i = 1, \ldots, m$). Now, one thing is left, that is to select one point within this area, for instance, the centre of gravity, and the problem to determine the function $W(\cdot)$ is solved. Now it has to be minimized by one of the methods considered above.

However, it is not always easy for an expert to give the opinion in such conditions. Being presented with arbitrary situations, the expert is going more often than not to refuse to make a decision than take it. These are the cases when the adaptive method should be applied.

The *adaptive method* is grounded on the use of the experts' decisions in the process of solving the multicriterial problem. Its essence may be reduced to the following. An expert is presented with an i-th situation $\mathbf{X}_i \in \Omega^*$ (within the Pareto set) and with the numerical values of all the criteria for this situations $q_1(\mathbf{X}_i)$, .. ., $q_m(\mathbf{X}_i)$. The expert has to make a decision along which ones ot the m criteria the next minimization step should be taken (to fhe detriment of the others, naturally). The procedure may be performed formally by passing over to the point $\mathbf{X}_{i+1} \in S^*$ with the properties

$$q_{k_j}(\mathbf{X}_{i+1}) < q_{k_j}(\mathbf{X}_i), \qquad \text{(S.6.6)}$$

where k_j ($j = 1, \ldots, k^i < m$) are the numbers of the criteria pointed out by the expert at the i-th step (all in all, k^i steps). This is namely the information I obtained at the i-th step. It is not difficult to solve problem (S.6.6), the more so because the point $\mathbf{X}_{i+1}$ should be close to $\mathbf{X}_i$. All the values of the criteria are determined for the obtained point $\mathbf{X}_{i+1}$, again to be presented to the expert, and so on until the expert stops at a solution $\mathbf{X}^*$ recognized as optimal. This is going to be the solution of the multicriterial problem.

Comments on References to Supplement

The optimization problem defined in § S.1 is in its essence a problem of mathematical programming with the difference that the minimized function and the functions of constraints may be specified as an algorithm, i.e. a rule to calculate their values, and not in an explicit analytical form, as in mathematical programming. This does not allow for using the data on the structure of the said functions in the process of seeking the solution, which makes us apply the universal enough seeking algorithms.

The given classification is very conventional (as are all classifications) and its aim is only to orient the reader in the structure of

the Supplement. A more comprehensive review can be found in the book [1].

To make the first acquaintance with the methods of seeking optimization, the reader is recommended the popular-science book [2].

The unidimensional search considered in § S.2 is not an independent algorithm here since very few problems of control may be reduced to optimization by one variable. It comes as a procedure of descent along the selected direction in the multiparametrical optimization. There are many methods of unidimensional optimization. Most comprehensively they are presented in [3].

§ S.3 deals mainly with the algorithms of descent for the multiparametrical optimization. However, it covers only an insignificant portion of the existing methods. Their detailed description may be found in [1-8]. Apart from that, to solve the optimization problems, one may apply some methods of non-linear programming not very much related to the nature of non-linearity (see references to Chapter 6).

The search for the global extremum considered in § S.4 seems to be the most involved optimization problem. Diverse algorithms for the random search are applied to solve it [1, 4, 6, 7, 9-11]. There are also determined algorithms for the global search [12, 13].

The idea of smoothing with the use of the Monte Carlo method provided a possibility to construe very efficient methods for the global search [14, 15].

The commonest method to solve the problems of conditional optimization is the method of fining functions described in § S.5. More details are available in [8, 16].

The problems of multicriterial vector optimization considered in § S.6 are rather a new class of problems, and the main problem here is the method of convoluting all the criteria into a global one. Various methods of convolution are discussed in [17-19]. A popular presentation of the problem to make optimal decisions in a multicriterial situation can be found in the brochure [20]. The reviews [21, 22] and the collection [23] may help to acquire full comprehension of the existing methods of multicriterial optimization.

References

1. Rastrigin, L.A. *Systems of Extremal Control.* Nauka, Moscow, 1974 (in Russian).
2. Pervozvanskii, A.A. *The Search.* Nauka, Moscow, 1972 (in Russian).
3. Wild, D.J. *Optimum Seeking Methods.* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1964.
4. Rastrigin, L.A. *Statistical Methods of Search.* Nauka, Moscow, 1968 (in Russian).
5. Pshenichnyi, B.N., Danilin, Yu.M. *Numerical Methods in Extremal Problems.* Nauka, Moscow, 1975 (in Russian).
6. Gurin, L.S., Dymarskii, Ya.S., Merkulov, A.D. *Problems and Methods of Optimal Distribution of Resources.* Sov. radio, Moscow, 1968 (in Russian).

7. Polovinkin, A.I. (ed.) *Algorithms of Optimization of Design Solutions.* Energiya, Moscow, 1978 (in Russian).
8. Himmelblau, D.M. *Applied Nonlinear Programming.* McGraw-Hill Book Co., New York, London, 1972.
9. Motskus, I.B. *Multiextremal Problems in Design.* Nauka, Moscow, 1967 (in Russian).
10. Yudin, D.B. *Mathematical Methods of Control in Conditions of Incomplete Information.* Sov. radio, Moscow, 1974 (in Russian).
11. Kagan, B.M., Ter-Mikaelyan, T.M. *Solution of Engineering Problems with Computers.* Energiya, Moscow, 1964 (in Russian).
12. Ulanov, G.M., Nosullaev, N.N. Multiextremal Systems of Control. In: *Technical Cybernetics, VINITI,* Moscow, 1972, v. 4 (in Russian).
13. Strongin R.G. Informational method of multiextremal optimization for measurements with noise. *Izv. Akad. Nauk SSSR, Tekhnicheskaya kibernetika,* No. 3 (1966).
14. Katkovnik, V.Ya. Methods of algorithmical optimization. In: Nelepin, R.A. (ed.) *Methods of Study of Nonlinear Systems of Automatic Control.* Nauka, Moscow, 1975 (in Russian).
15. Katkovnik, V.Ya. *Linear Estimates and Stochastic Problems of Optimization: The Method of Parametrical Operators of Averaging.* Nauka, Moscow, 1976 (in Russian).
16. Fiacco, A.V., McCormic, G.P. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* Wiley, Inc., New York, 1968.
17. Germeyer, Yu.B. *Introduction into the Theory of the Study of Operations.* Nauka, Moscow, 1971 (in Russian).
18. Salukvadze, M.E. *Problems of Vector Optimization in the Theory of Control.* Metsniereba, Tbilisi, 1975 (in Russian).
19. Gutkin, L.S. *Optimization of Radioelectronic Devices by the Set of Indices of Quality.* Sov. radio, Moscow, 1975 (in Russian).
20. Solnyshkov, Yu.S. *Quantitative Substantiation of Decisions.* Znanie, Moscow, 1971 (in Russian).
21. Larichev, O.I. Man-machine procedures of making decisions. *Avtomatika i telemekhanika,* No. 12 (1971).
22. Emel'anov, S.V., Borisov, V.I., Molevich, A.A., Cherkashin, A.M. Models and methods of vector optimization. In: *Results of Science and Engineering. VINITI,* Moscow, *Tekhnicheskaya kibernetika,* No. 5, 1973.
23. Gvishiani, D.M., Emel'anov, S.V. (eds.) *Multicriterial Problems of Making Decisions.* Mashinostroeniye, Moscow, 1978 (in Russian).

Subject Index

Absolutely controllable object 70-71
Active object 33, 201-203
Adaptation of model 34
Adaptive algorithm of identification 116-118, Fig. 4.1.1
Adaptive method 244
Ageing of data 221
Agreement of experts 88
Algorithm of control 20-21, Fig. I.2.1
A priori and a posteriori data 115

Binary relations language 107-112
'Black box' 80, 94
Boundaries of object 60-63, Fig. 2.1.1

Competence quotients 64
Complex objects 22-25
  'intolerance' to control 24
  irreproducibility of experiments 24
  non-stationarity 24
  stochastic behaviour 23-24
Complexity of model 95-99
Compromise area 240
Control 5-7, 8-9, 13-18, 21
  algorithm 20
  complex object of 22-25
  formal approach to 8-9
  informal approach to 9
  object of 13
  of dynamic object 161
  optimal 13, 160, 192
  realization of 156-162, 198-204
  stages of 25-36, Fig. I.4.1
  subject of 18
  synthesis of 156-162
  system 18-22, Figs. I.2.1, 1.2.1
Controllability of object 67-69
  expert estimation of 69-70, 75-78
  partial or relative 71
Contrast matrix 92-93, Fig. 3.2.4
Convex programming 163-165
Correction 35-36, 206-223
Criteria for experiment design 134, 137-139, Fig. 5.1.1
Criterion of experts' agreement 88

Design resources 134, Fig. 5.1.1
Discomfort function 49
Discriminating experiment design 149-154
Dispersion matrix of parameters 137-138
D-optimal experiment design 139-146
Drift of situations 57, Figs. 1.3.1, 1.3.3
Dynamic programming 186-195

Efficiency indices for structures 213
Equivalence 49
Experiment design 30-33, 131-155
  criteria for 134, 137-139, Fig. 5.1.1
Experiment
  discriminating 149-154
  D-optimal 139-146
  step-by-step 146-148
Expert evaluation 63-67, 75-78
Experts' judgement of uncontrollability of object 78

Finite difference model 105
Fisher information matrix 122
Forgetting function 213-215
Formal approach to control 8-9

Hamiltonian function 183-184

Identification
  adaptive algorithm of 116-118, Fig. 4.1.1
  of dynamic objects 125-129
  of parameters 29-30, 114-129
  of static objects 118-125
Informal approach to control 9
Inputs and outputs 18-20, 80-83, Fig. I.2.1
  expert ranking of 83-94.
Integer programming 166
Interaction between subject and control system 49, Fig. 1.2.1
Isolating object from medium 60
'Intolerance' to control 24

Mathematical programming.
  analysis of problems 169-173
  classification of problems 162-165
  convex 163-165
  dynamic 186-195
  integer 166
  linear 162-163
  non-linear 165-166
  stochastic 166-168
Matrix
  contradiction 91, Fig. 3.2.3
  contrast 92-93, Fig. 3.2.4
  design 133
  dispersion (of parameters) 137-138
  Fisher information 122
  of competence evaluation 64-65
  of mean preference 92, Fig. 3.2.4
Maximum principle 182-186, Fig. 6.5.1
Method
  a priori 242
  a posteriori 243
  coordinate (Gauss-Seidel) 229
  dichotomy 226-227
  direct ranking 85-90
  direct search 180-181
  global random search 231
  golden section 227-228

Method
gradient 229-230
multidimensional stochastic approximation 230
of fining functions 236-239
of global random search 231
of pair comparisons 90-94
of possible directions 176-178
of reverse search 181
of smoothing global search 231-233
of tangents 225-226
projection 235-236
recurrent 169
simplex 173-176
substitution 233-235
Multi-port 94, Fig. 3.3.2

Negotiation set 240
Non-linearity of model 101-103
Non-stationarity 24, 103-104

Objective of control 14-16, 55
Objective variables 53
Object of control 13
Operator 14
Optimal control 13, 160, 192
Optimal objective 51
Optimization problems, characteristics of 224
Optimum of 'size' of object 61-63

Pair comparisons 90-94
Paradox of objective 46-47
Pareto set 240
Passive observation 30
Preference 49

Quality functions 171-172, Figs. 6.3.2, 6.3.3

Ranking 83-94
Realization of control 33, 156-162, 198-204
Resources of design 134, Fig. 5.1.1

Self-adjusting model 117, Fig. 8.1.1
Semiotic models 106-112
Simulation models 104-106
Stochastic behaviour 23-24
Stochasticity of model 102-103
Structural adjustment 223
Structural synthesis of model 28-29
Structure of optimum seeking problems 224, Fig. S.1.1
Subject of control 18
Syntactically correct phrases 108
Systems approach 44

Transformation of initial description 111-112
Transitivity rule 92

Unobserved disturbance 19, 82, Fig. 1.2.1

Printed in the Union of Soviet Socialist Republics